Report G2M Case Study

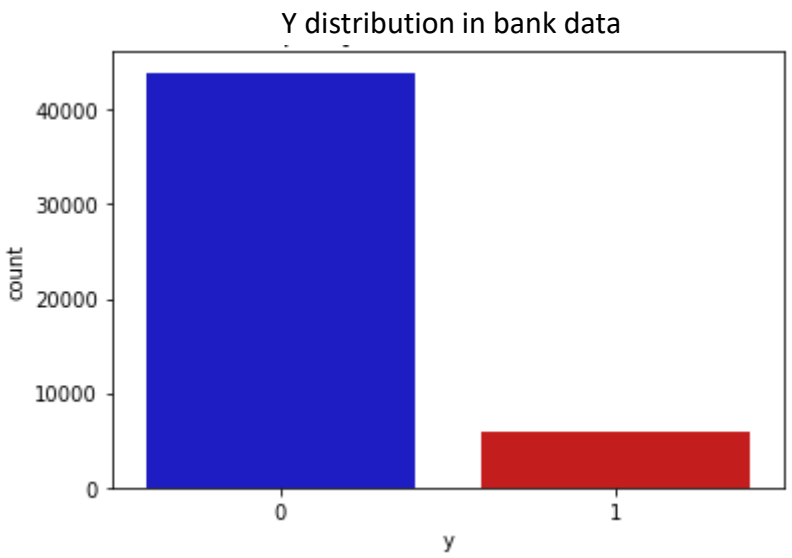
Virtual Internship

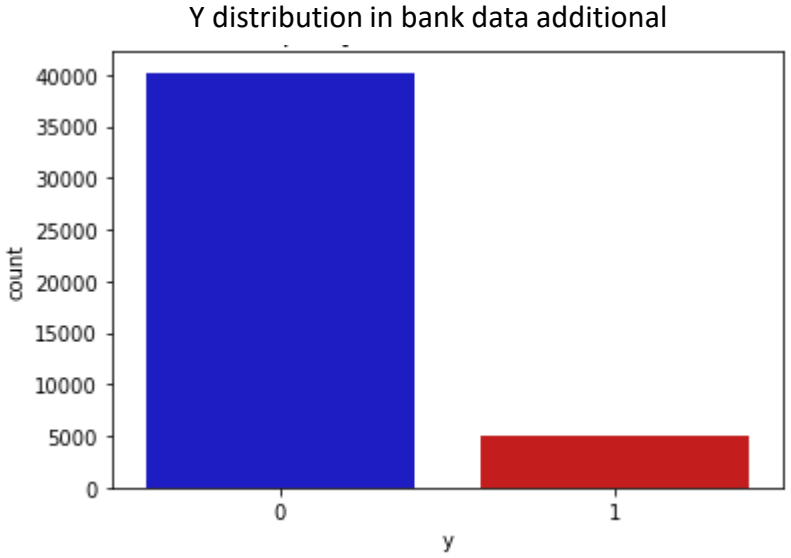
28-Jan-2024

Background –G2M(bank marketing) case study

- ABC Bank wants to sell its term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).
- Objective : Bank wants to use ML model to shortlist customer whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc) can focus only to those customers whose chances of buying the product is more.
- Recommended models : Logistic Regression, K Nears Neighbor, Support Vector Classifier and Decision Tree Classifier

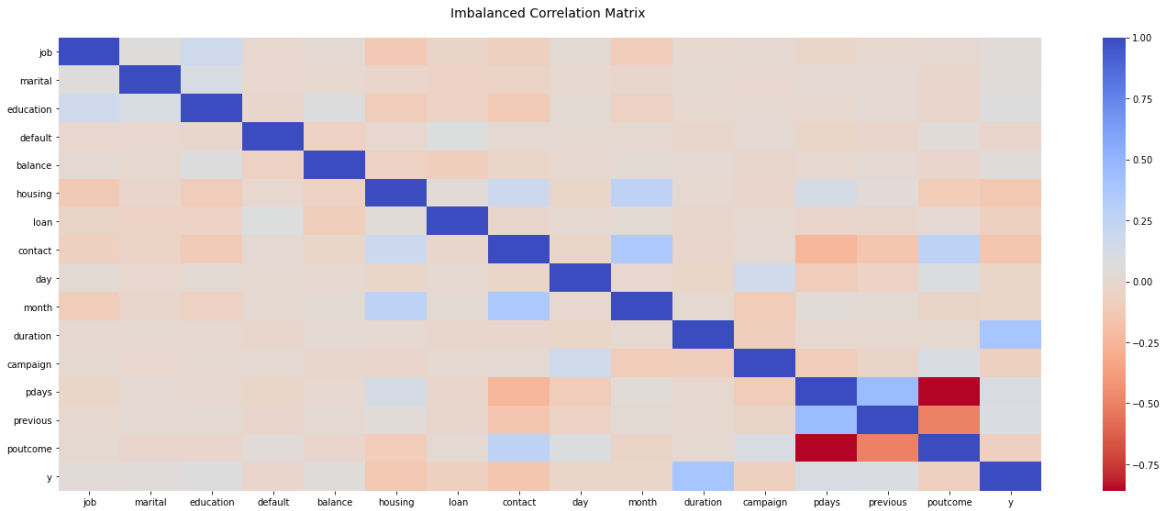
Y Distribution in both datasets

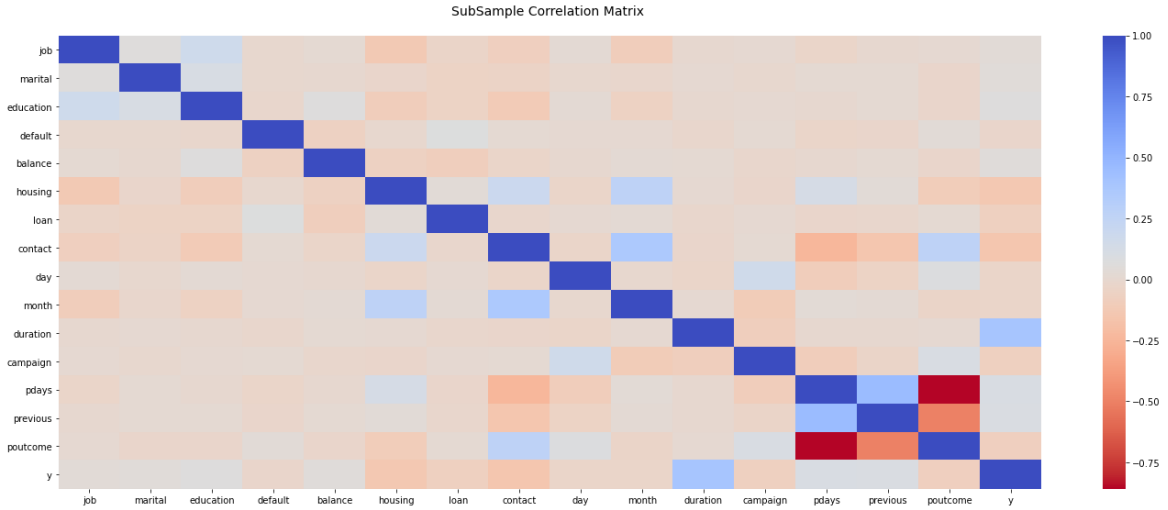




- The distribution of the Y feature in both datasets are very similar, with more rows in the original bank data

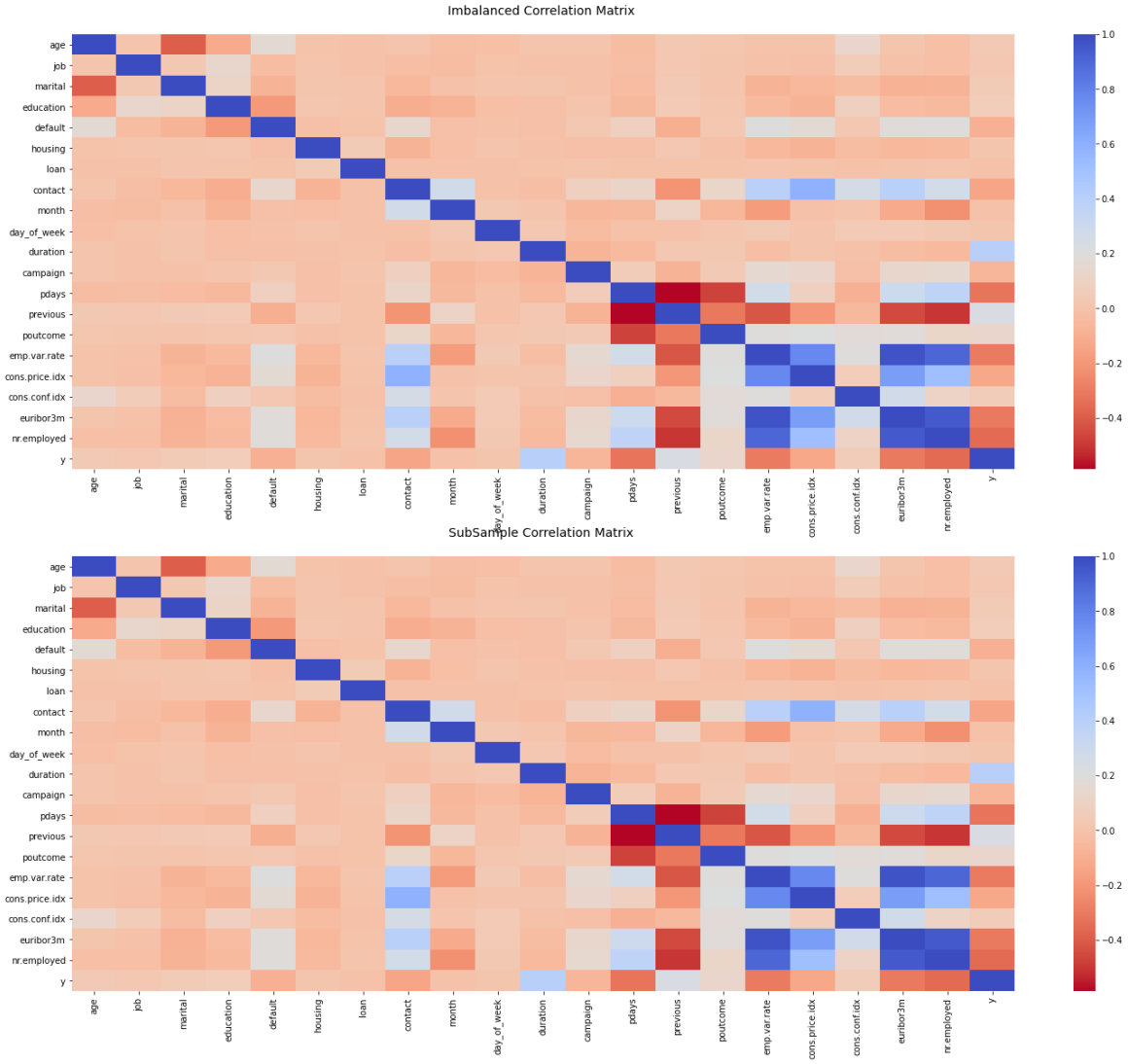
Correlation Matrix for Bank Data Dataset





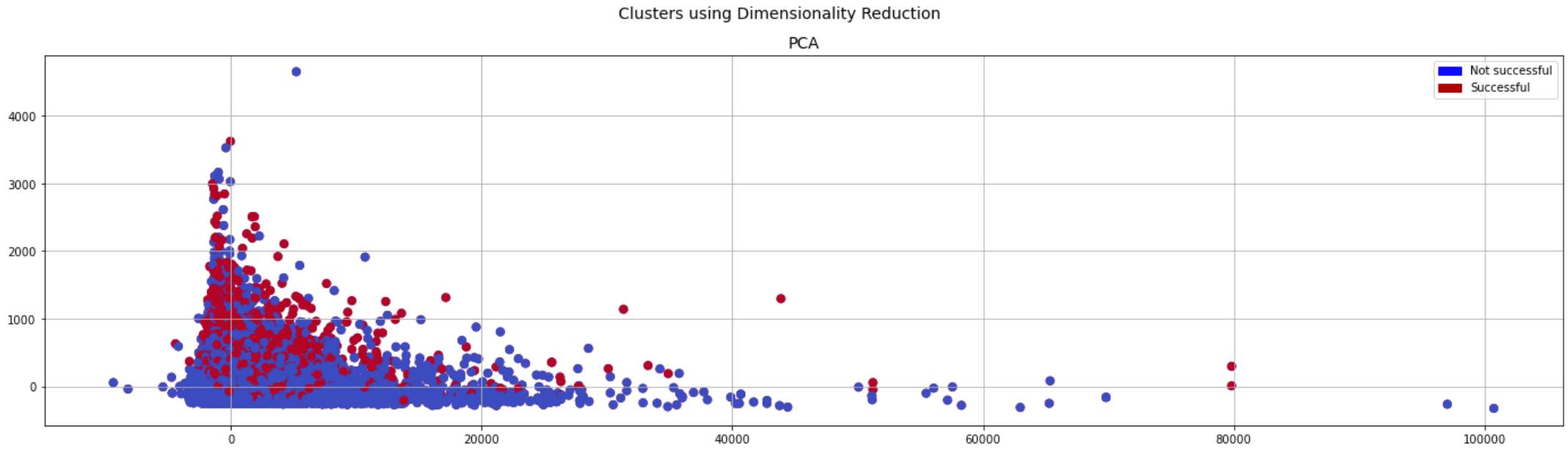
- The correlations between the features are not very strong with little difference between the imbalanced and subsample correlation matrix

Correlation Matrix for Bank Data Additional Dataset



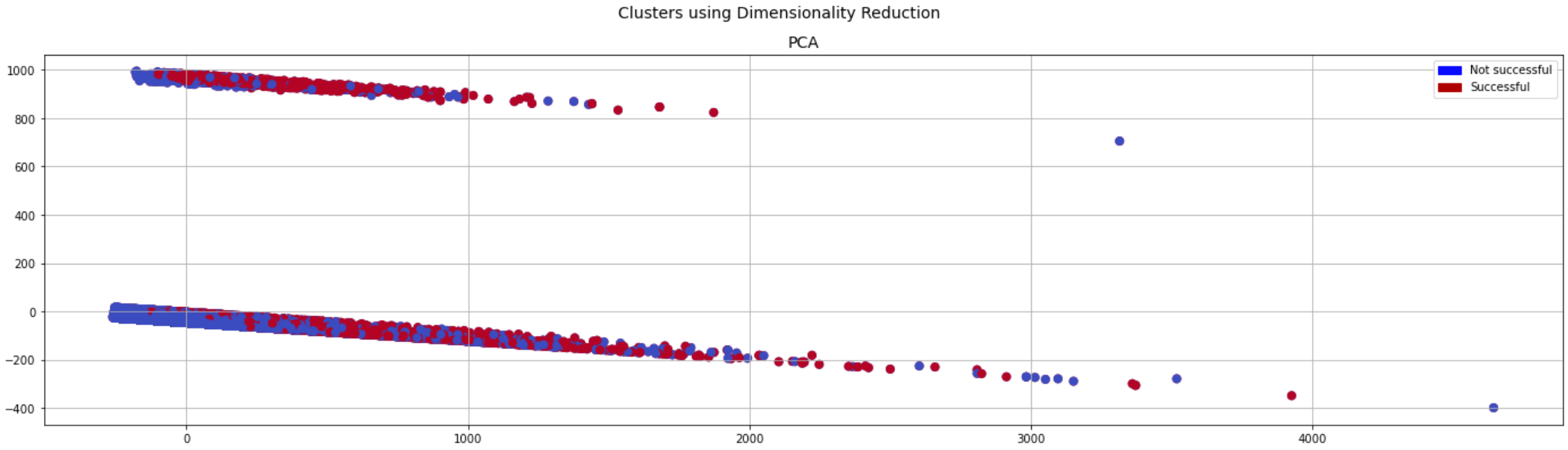
- The correlations between the features are not very strong with little difference between the imbalanced and subsample correlation matrix
- However, in the bank additional dataset, the correlation between the features are weaker than in the original bank dataset

PCA Clusters for the Bank Dataset



- As seen here, there are no cluster group which could differentiate between the two group in the classification problem

PCA Clusters for the Bank Additional Dataset



- Like the original dataset, the two group could not be differentiated

Training Score for both Datasets

- Bank Data Dataset
- Logistic Regression Has a training score of 89.0 % accuracy score
- K neighbors Classifier has a training score of 88.0 % accuracy score
- Decision Tree Classifier has a training score of 89.0 % accuracy score

- Bank Additional Dataset
- Logistic Regression Has a training score of 91.0 % accuracy score
- K neighbors Classifier has a training score of 91.0 % accuracy score
- Decision Tree Classifier has a training score of 90.0 % accuracy score

- As seen here, bank additional dataset has a better training accuracy score in all three models than the original bank dataset

Cross Validation Score for both Datasets

- Bank Data Dataset
- Logistic Regression Cross Validation Score: 88.7%
- K nears Neighbors Cross Validation Score 88.49%
- Decision Tree Classifier Cross Validation Score 89.06%

- Bank Additional Dataset
- Logistic Regression Cross Validation Score: 90.89%
- K nears Neighbors Cross Validation Score 90.67%
- Decision Tree Classifier Cross Validation Score 90.83%

- As seen here, bank additional dataset has a better cross validation score in all three models than the original bank dataset

ROC AUC score for both Datasets

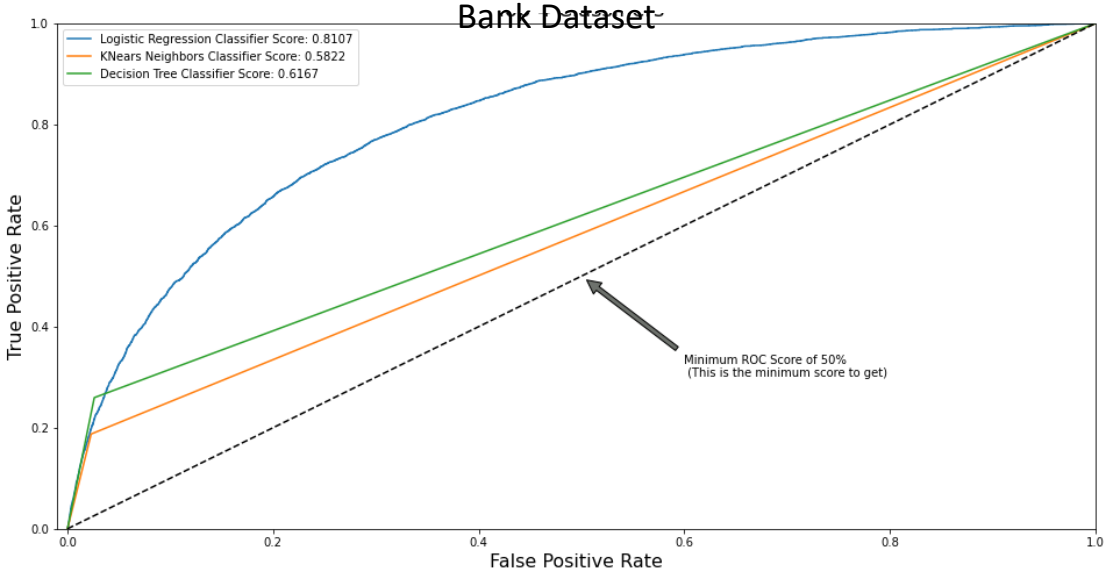
- Bank Data Dataset
- Logistic Regression ROC AUC score: 0.8107350754432993
- K Nears Neighbors ROC AUC score: 0.5822402117004325
- Decision Tree Classifier ROC AUC score: 0.6167092827931584

- Bank Additional Dataset
- Logistic Regression ROC AUC score: 0.924293322821365
- K Nears Neighbors ROC AUC score: 0.6811811165228423
- Decision Tree Classifier ROC AUC score: 0.7670421453264366

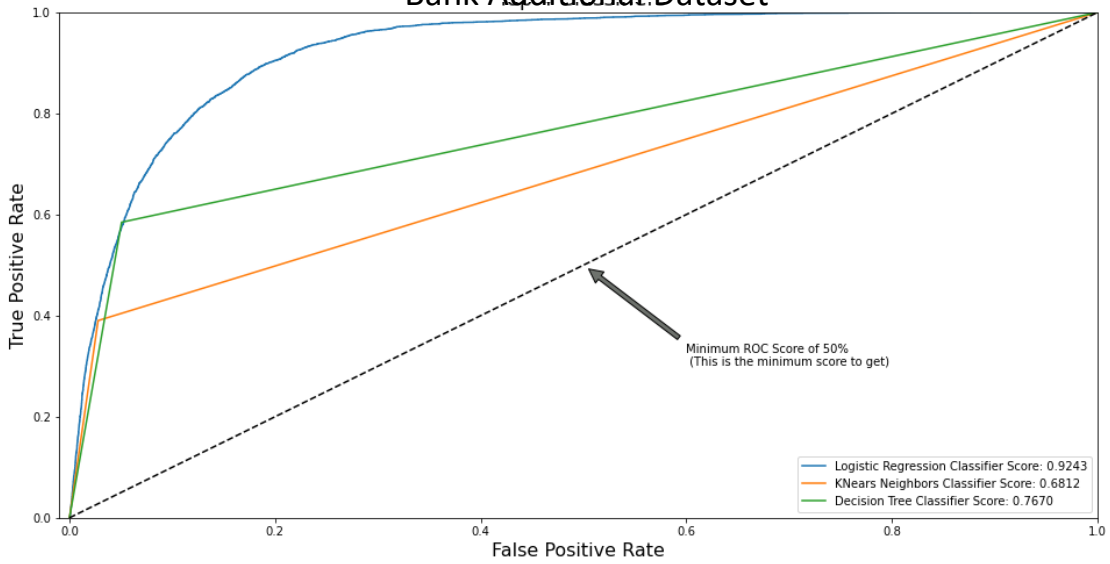
- As seen here, bank additional dataset has a better ROC AUC score in all three models than the original bank dataset

ROC Curve in both Datasets

ROC Curve Top 3 Classifiers
Bank Dataset



ROC Curve Top 3 Classifiers
Bank Additional Dataset



- In both datasets, logistic regression classifier has a better ROC curve than the other two classifier
- However, in the bank additional dataset, the logistic regression classier has a better score. This means that the bank additional dataset has a better true positive rate than the original bank dataset

SMOTE Prediction Score for both Datasets

SMOTE Prediction Score for Bank Dataset

	precision	recall	f1-score	support
"no"	0.95	0.82	0.88	8732
"yes"	0.35	0.70	0.46	1215
accuracy			0.80	9947
macro avg	0.65	0.76	0.67	9947
weighted avg	0.88	0.80	0.83	9947

SMOTE Prediction Score for Bank Additional Dataset

	precision	recall	f1-score	support
"no"	0.98	0.85	0.91	8031
"yes"	0.42	0.70	0.46	1031
accuracy			0.85	9062
macro avg	0.70	0.85	0.73	9062
weighted avg	0.91	0.85	0.87	9062

- SMOTE refers to a method of balancing the dataset between the two classification groups by creating new data for the smaller group
- Based on these tables, the bank additional dataset has better SMOTE Prediction Scores than the original bank dataset
- The positive group has lower scores than the negative group in both datasets

Logistic Regression Prediction Scores for both Datasets

Logistic Regression Prediction Score for Bank Dataset

Logistic Regression				
	precision	recall	f1-score	support
0	0.96	0.83	0.89	8777
1	0.35	0.71	0.47	1170
accuracy			0.81	9947
macro avg	0.65	0.77	0.68	9947
weighted avg	0.88	0.81	0.84	9947

Logistic Regression Prediction Score for Bank Additional Dataset

Logistic Regression				
	precision	recall	f1-score	support
0	0.98	0.85	0.91	8046
1	0.42	0.86	0.56	1016
accuracy			0.85	9062
macro avg	0.70	0.85	0.74	9062
weighted avg	0.92	0.85	0.87	9062

- As seen in those tables, the Bank Additional Dataset has better scores than the original dataset

K Nears Neighbors Prediction Scores for both Datasets

K Nears Neighbor Prediction Score for Bank Dataset

K Nears Neighbors				
	precision	recall	f1-score	support
0	0.90	0.98	0.94	8777
1	0.54	0.19	0.28	1170
accuracy			0.89	9947
macro avg	0.72	0.59	0.61	9947
weighted avg	0.86	0.89	0.86	9947

K Nears Neighbor Prediction Score for Bank Additional Dataset

K Nears Neighbors				
	precision	recall	f1-score	support
0	0.93	0.97	0.95	8046
1	0.66	0.40	0.50	1016
accuracy			0.91	9062
macro avg	0.79	0.69	0.73	9062
weighted avg	0.90	0.91	0.90	9062

- As seen in those tables, the Bank Additional Dataset has better scores than the original dataset

Decision Tree Classifier Prediction Scores for both Datasets

Decision Tree Classifier Prediction Score for Bank Dataset

Decision Tree Classifier				
	precision	recall	f1-score	support
0	0.91	0.98	0.94	8777
1	0.60	0.26	0.36	1170
accuracy			0.89	9947
macro avg	0.75	0.62	0.65	9947
weighted avg	0.87	0.89	0.87	9947

Decision Tree Classifier Prediction Score for Bank Additional Dataset

Decision Tree Classifier				
	precision	recall	f1-score	support
0	0.95	0.95	0.95	8046
1	0.59	0.60	0.59	1016
accuracy			0.91	9062
macro avg	0.77	0.77	0.77	9062
weighted avg	0.91	0.91	0.91	9062

- As seen in those tables, the original bank dataset has better scores than the original dataset

Conclusion

- Overall, the bank additional dataset has better prediction score in all three models. However, made focusing on the two classification groups, the negative group has better prediction score in all models in both datasets than the positive group.
- The best all-around model is the logistical regression. It could be used in predicting negative outcomes in both datasets. As for positive outcomes, K Near Neighbor had the best precision score for bank additional dataset, while at the same time Decision Tree Classifier has the best precision score
- In conclusion, the Logistical Regression is the best model suited for both datasets and it is recommended for solving the classification problem of this case study

Thank You