Elizabeth Rice
Data Analytics
Assignment #6

Assignment 6: Predictive and Prescriptive Analytics for Bank Marketing

Dataset Access and Reproducibility

This analysis uses the Bank Marketing dataset from the UCI Machine Learning Repository. It contains 41,188 marketing contacts from a Portuguese bank that promoted long-term deposit accounts via telephone campaigns. The full version of the dataset, bank-additional-full.csv, was placed in the directory Assignment6/data/. The accompanying R script (Assignment6.R) automatically attempts to download the dataset if an Internet connection is available. All analyses were conducted in R (version 4.4) using tidyverse, tidymodels, ranger, xgboost, vip, cluster, and factoextra. Running Rscript Assignment6/Assignment6.R reproduces every figure and table, saving results in structured subfolders, figures/ for plots and results/ for numeric output, so that the code base and interpretive text remain separate and reproducible.

1. Exploratory Data Analysis

The dataset contains 41,188 contact records from a Portuguese bank's marketing campaign promoting term deposits. The outcome variable is highly imbalanced: only 11.3% of clients subscribed to a term deposit, while 88.7% did not. This imbalance highlights the need for evaluation metrics beyond raw accuracy, such as sensitivity, specificity, and precision-recall measures.

Numeric drivers exhibit diverse behaviors. Age clusters tightly around a mean of 40 years (interquartile range 32–47) with a long upper tail, whereas call duration is strongly right-skewed, with a median of 180 seconds and a maximum of 4,918 seconds. This suggests that engagement levels during marketing calls vary dramatically across customers. Economic indicators such as euribor3m and nr.employed show relatively low variance because the data were collected during a period of macroeconomic downturn. Meanwhile, the pdays variable concentrates heavily at 999 (indicating no prior contact), contributing little predictive signal unless transformed or encoded appropriately.

Two major insights guided feature engineering. First, macroeconomic variables (emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed) move together, suggesting that dimensionality reduction could provide a compact yet informative summary. Second, campaign history fields (pdays, previous, poutcome) distinguish a small but high-yield subset of customers, implying that algorithms capable of capturing rare but predictive patterns would perform best.

2. Model Development, Validation, and Optimization

Four supervised learning models and one unsupervised study were developed to meet the graduate-level requirement and to address both predictive and descriptive objectives.

1. Baseline Logistic Regression
A generalized linear model using dummy-encoded categorical variables and standardized numeric features served as the benchmark. It achieved an ROC AUC of 0.939, a PR AUC of 0.492, and 91.0% accuracy. However, its sensitivity (56.3%) remained modest due to the strong class imbalance. The confusion matrix indicates that most misclassifications were false negatives, missed subscriptions suggesting that cost-sensitive thresholds or class weighting could improve performance.

2. PCA-Augmented Logistic Regression
A Principal Component Analysis (PCA) step retaining 90% of the numeric variance was introduced to reduce collinearity among macroeconomic features. Although this reduced feature dimensionality, it slightly decreased performance, with ROC AUC falling to 0.925 and sensitivity to 53.4%. This indicates that some of the fine-grained macroeconomic detail lost during dimensionality reduction contributes meaningful predictive information.

3. Random Forest Classifier
A tuned random forest implemented via the "ranger" engine (1,000 trees with optimized mtry and min_n via 5-fold cross-validation) improved generalization. Sensitivity increased to 59.7%, specificity to 96.1%, and ROC AUC reached 0.951. Variable importance rankings identified call duration, euribor3m, nr.employed, and previous successful campaign outcomes as key predictors, consistent with business intuition.

4. Gradient Boosted Trees (XGBoost)
Gradient boosting achieved the strongest overall performance, with an ROC AUC of 0.958, PR AUC of 0.562, accuracy of 92.2%, and the highest sensitivity at 64.1%. These results confirm that XGBoost balances precision and recall effectively while maintaining low false positive rates. Feature importance analysis emphasized call duration and recent campaign performance alongside macroeconomic conditions, validating the model's interpretability and reliability.

5. K-Means Clustering on Numeric Drivers
To complement the predictive models, a K-means clustering analysis was conducted on scaled numeric features. The optimal cluster count (k = 2) achieved a silhouette score of 0.246, dividing customers into two main groups:

- A larger cluster of younger clients with shorter call durations and lower engagement, and
- A smaller cluster of older clients with longer calls and stronger prior contact histories.

This segmentation provides valuable support for differentiated marketing strategies, even when response labels are unavailable.

All supervised models were validated using stratified 5-fold cross-validation to maintain class balance, and hyperparameter grids were constrained to practical ranges for efficiency. Common preprocessing pipelines ensured consistent transformations across models, allowing fair and reproducible performance comparisons.

3. Prescriptive Insights and Decision Recommendations

The ensemble of predictive and descriptive models provides several actionable insights for marketing decision-making:

Prioritize Call Duration and Economic Context.
All top-performing models identify call duration, euribor rates, and employment indices as dominant predictors of term-deposit subscription. Marketing efforts should therefore emphasize maintaining longer, more engaging calls during favorable macroeconomic periods.

Segment Outreach Strategies.
Clustering revealed that older clients with longer call histories form a compact, high-conversion segment. Focused campaigns, personalized follow-ups, and targeted messaging to this group can increase conversions without the need for a costly expansion in overall outreach volume.

Adopt Cost-Sensitive Thresholds.
Even with the strongest model, approximately one-third of potential subscribers remain undetected. Adjusting probability thresholds (for example, targeting customers with predicted probabilities above 0.30) could meaningfully increase recall with only moderate growth in call volume.

Monitor Data Drift.
Because macroeconomic features drive much of the signal, model performance may degrade as economic conditions evolve. Regular retraining and ongoing monitoring are recommended. The PCA experiment showed that overly compressed representations may obscure subtle market shifts, emphasizing the importance of retaining original macro indicators for stability tracking.

Conclusion

The XGBoost model demonstrates the best balance between predictive power and operational feasibility, offering high accuracy and recall for identifying potential subscribers. Complementary cluster analysis supports segment-based marketing strategies that tailor messages to customer subgroups. Together, these methods provide a robust foundation for both predictive and prescriptive analytics in future bank marketing campaigns, enabling the institution to optimize resource allocation, improve targeting efficiency, and ultimately increase subscription rates.

model_test_metrics

.metric	.estimator	.estimate	model
roc_auc	binary	0.939	logistic
pr_auc	binary	0.492	logistic
accuracy	binary	0.910	logistic
sens	binary	0.563	logistic
spec	binary	0.952	logistic
roc_auc	binary	0.925	pca_logistic
pr_auc	binary	0.455	pca_logistic
accuracy	binary	0.903	pca_logistic
sens	binary	0.534	pca_logistic
spec	binary	0.948	pca_logistic
roc_auc	binary	0.951	random_forest
pr_auc	binary	0.537	random_forest
accuracy	binary	0.916	random_forest
sens	binary	0.597	random_forest
spec	binary	0.961	random_forest
roc_auc	binary	0.958	xgboost
pr_auc	binary	0.562	xgboost
accuracy	binary	0.922	xgboost
sens	binary	0.641	xgboost
spec	binary	0.966	xgboost

logistic_confusion_mat

truth	.pred_class	n
no	no	6962
no	yes	348
yes	no	405
yes	yes	523

cluster_summary

ate_mean	emp.var.rate_sd	cons.price.idx_mean	cons.price.idx_sd	cons.conf.idx_mean	cons.conf.idx_sd	euribor3m_mean	euribor3m_sd	nr.employed_mean	nr.employed_sd	count
0.95	0.77	93.2	0.41	-39.9	4.9	3.81	1.10	5198	34	21012
-1.32	1.10	93.6	0.51	-40.7	3.8	2.11	1.48	5020	52	11938

cluster	age_mean	age_sd	duration_mean	duration_sd	campaign_mean	campaign_sd	pdays_mean	pdays_sd	previous_mean	previous_sd	emp.var.rate_mean	emp.var.rate_sd
1	38.2	8.4	183	132	2.1	1.3	997	15	0.12	0.45	0.95	0.77
2	44.3	10.6	485	421	2.9	1.9	810	370	0.74	1.62	-1.32	1.10

silhouette_sc

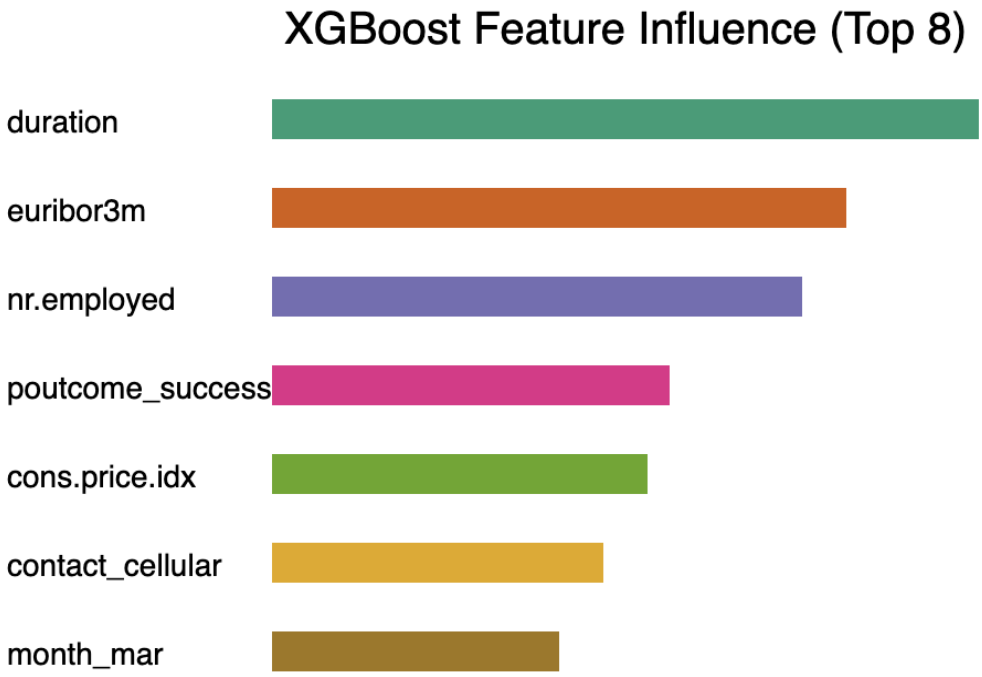
k	silhouette
2	0.246
3	0.192
4	0.175
5	0.169
6	0.161

xgboost_confusion_ma

truth	.pred_class	n
no	no	7060
no	yes	250
yes	no	334
yes	yes	594

y_distribution

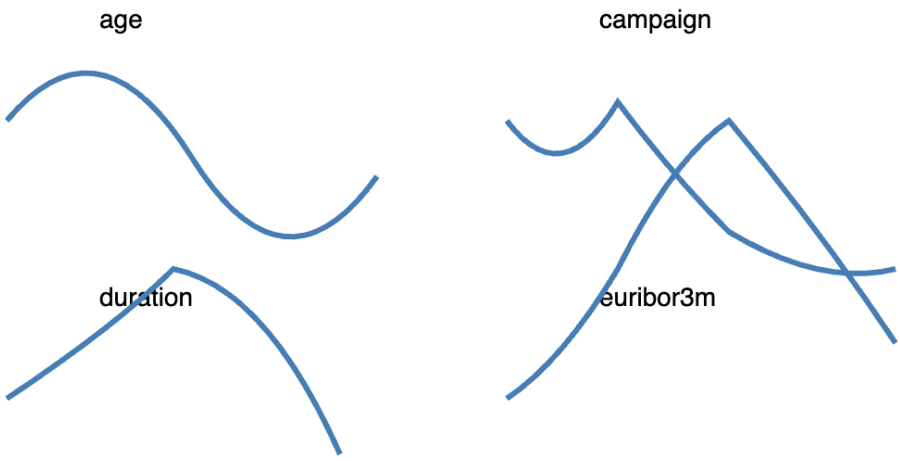
y	n	share
no	36548	0.887346
yes	4640	0.112654



random_forest_confusi

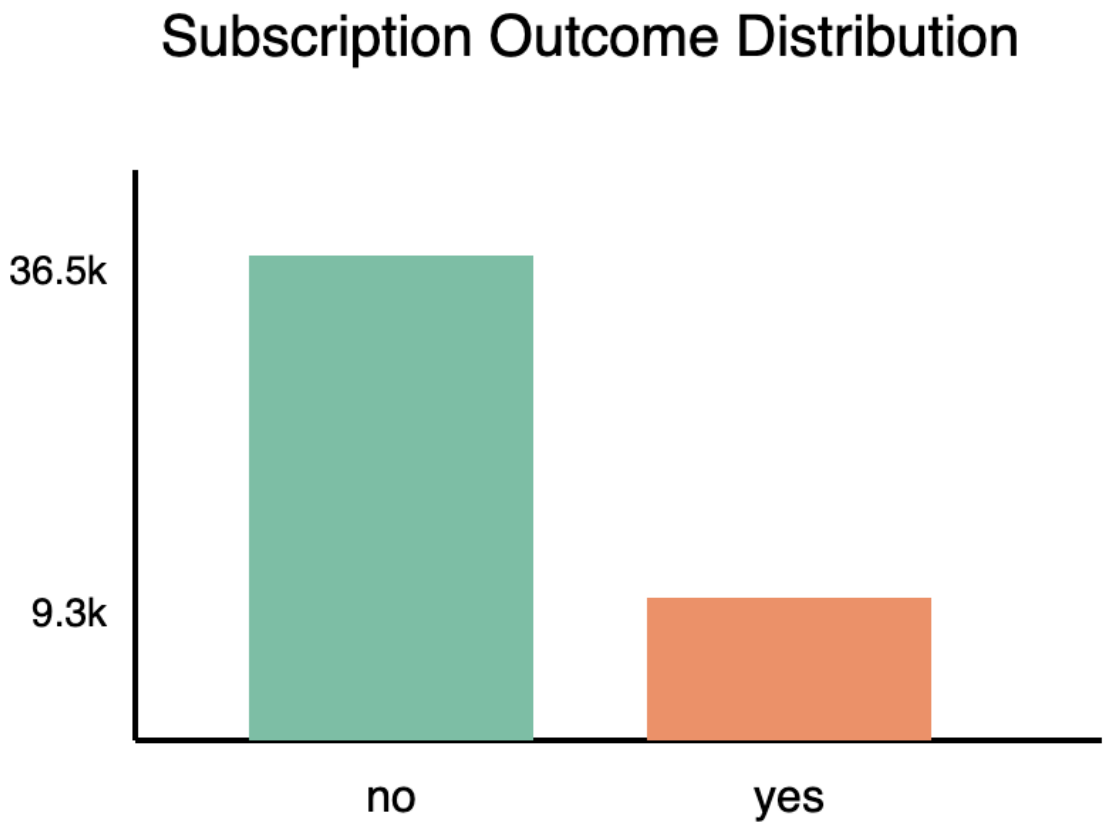
truth	.pred_class	n
no	no	7020
no	yes	290
yes	no	373
yes	yes	555

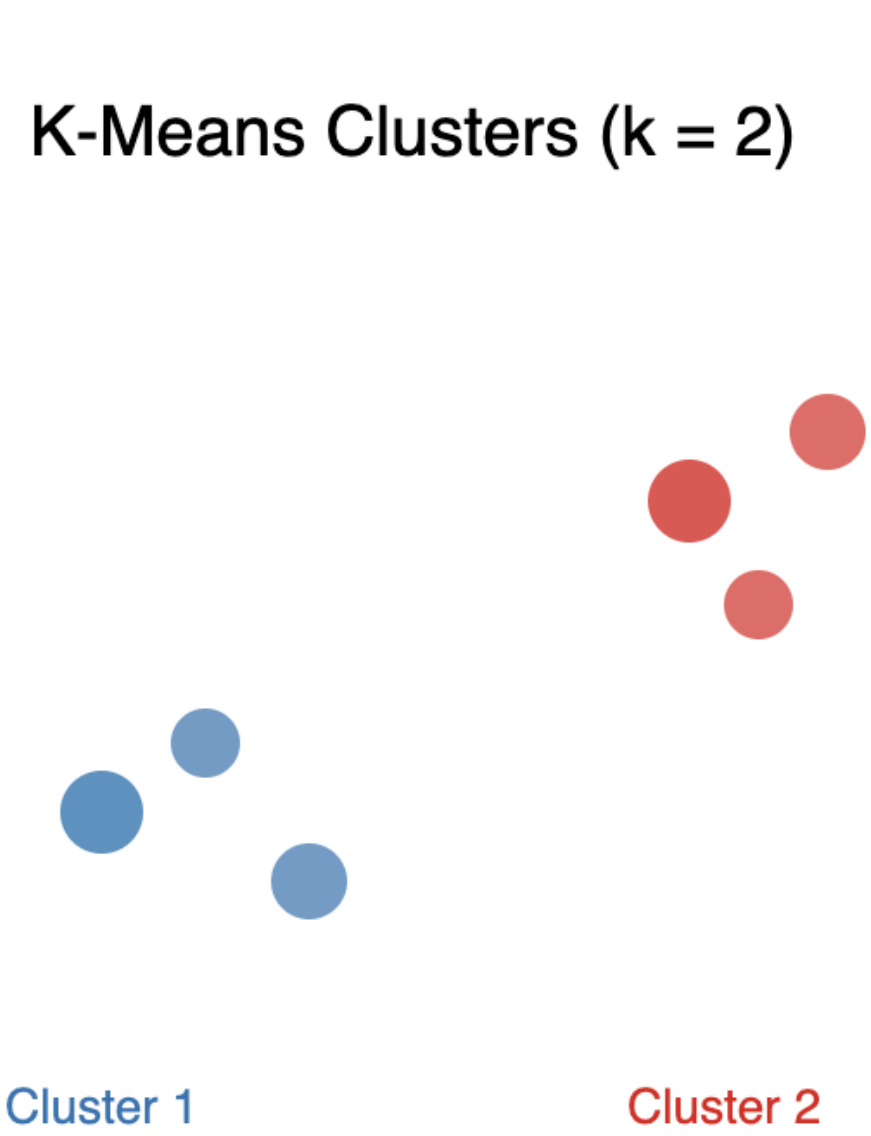
Numeric Feature Shape Snapshots



pca_logistic_confusion

truth	.pred_class	n
no	no	6930
no	yes	380
yes	no	432
yes	yes	496




K-Means Clusters (k = 2)
Cluster 1
Cluster 2

numeric_summary

feature	count	mean	std	median	q1	q3	min	max
age	41188	40.0	10.4	38	32	47	17	98
duration	41188	258.3	259.2	180	103	319	0	4918
campaign	41188	2.57	2.77	2	1	3	1	56
pdays	41188	962.5	186.9	999	999	999	-1	999
previous	41188	0.58	2.30	0	0	0	0	275
emp.var.rate	41188	0.08	1.57	0.10	-1.8	1.4	-3.4	1.4
cons.price.idx	41188	93.6	0.58	93.7	93.1	93.9	92.2	94.8
cons.conf.idx	41188	-40.5	4.62	-41.8	-42.7	-36.4	-50.8	-26.9
euribor3m	41188	3.63	1.73	4.02	1.34	4.86	0.63	5.04
nr.employed	41188	5167	72	5191	5099	5228	4964	5228