

Identifying High-Potential Customers Using Demographic Factors

Decision Tree Analysis with SMOTENC and SHAP

Project Context

This analysis investigates demographic factors associated with customers' likelihood of subscribing to a bank marketing campaign. The objective is to identify high-potential customer segments to support more precise targeting and campaign design.

Methods: Decision Tree (tuned), one-hot encoding, SMOTENC oversampling for class imbalance, permutation importance, and SHAP for interpretability.

Data source: UCI Machine Learning Repository – Bank Marketing Dataset (<https://archive.ics.uci.edu/ml/datasets/bank+marketing>).

1. Model Performance (Decision Tree)

Binary classification performance is reported using both ROC-AUC and PR-AUC. ROC-AUC reflects overall ranking ability across thresholds, while PR-AUC focuses on the positive class and is more informative under class imbalance.

Dataset note: The test-set positive (subscription) rate is 11.70%, so a random classifier would achieve PR-AUC ≈ 0.117 .

1.1 Baseline Tree

- Test PR-AUC = 0.210, about 1.8 \times the random baseline, indicating meaningful lift for identifying subscribers under class imbalance.
- Test ROC-AUC = 0.635 and Train ROC-AUC = 0.657, suggesting mild overfitting and moderate discrimination.

Overall, the baseline tree provides some signal but remains weak at capturing the positive class, which is the primary business interest.

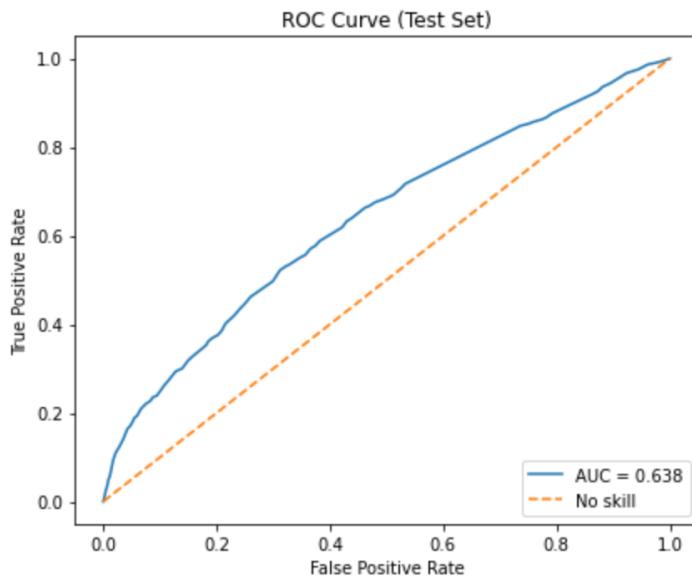
1.2 Tuned Tree

I tuned max_depth, min_samples_leaf, and min_samples_split via cross-validation. The best model achieved:

- Cross-validated ROC-AUC = 0.659

- Test ROC-AUC = 0.638; Train ROC-AUC = 0.666
- Best PR-AUC ≈ 0.212

Figure 1. ROC Curve



Test ROC-AUC: 0.638

ROC-AUC improves slightly while PR-AUC remains nearly unchanged. This suggests that with demographics-only features, the achievable lift is limited; substantial improvement likely requires richer financial and behavioural variables.

Implementation note: All preprocessing steps (one-hot encoding and SMOTENC) were applied within training folds to avoid data leakage.

2. Interpreting the Tree: Key Drivers and Segments

Although demographics alone do not yield high predictive accuracy, the decision tree provides clear and interpretable customer segmentation.

2.1 Key Splits (Top Levels)

- Age is the primary driver: the first split at approximately 60 separates groups with markedly different subscription rates.
- Marital status is a secondary driver: under 60, single vs. non-single matters; over 60, divorced vs. not divorced matters.
- Interactions matter: the impact of education and marital status depends on the age group.

Figure 2. Decision tree (top3 levels). Darker blue nodes indicate higher subscription probability.

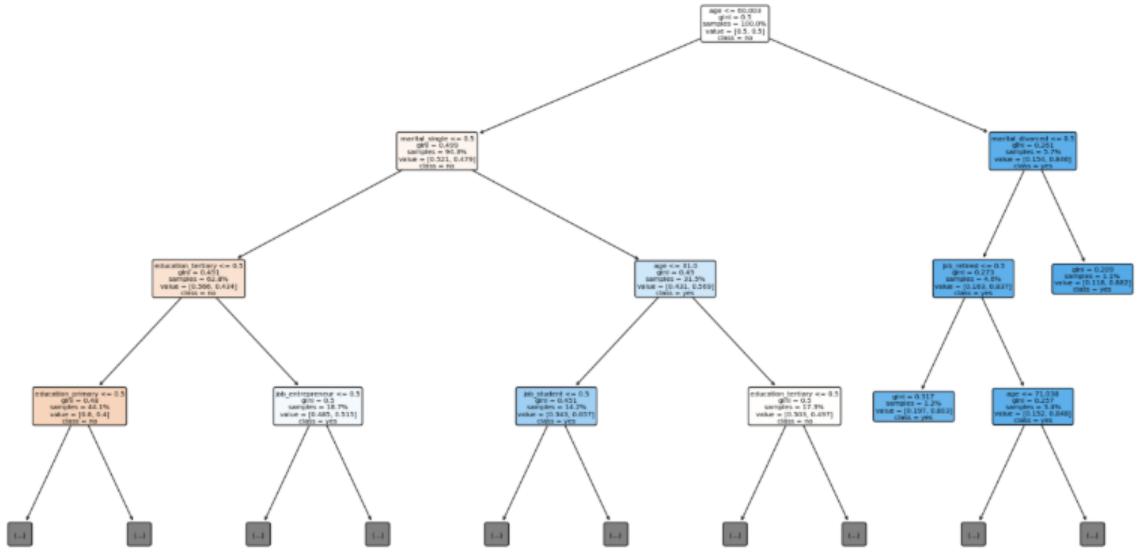


Table 2.1. Main splitting rules (top three levels).

Tree level	Feature	Split condition	Branch meaning	Interpretation
Level 1	age	age \leq 60	Left: age \leq 60; Right: age $>$ 60	Customers older than \sim 60 show higher predicted subscription probability.
Level 2	marital_single	marital_single = 1 vs 0 (within age \leq 60)	Single customers split from non-single	Among customers \leq 60, being single is associated with higher predicted probability (conditional on later splits).
Level 2	marital_divorced	marital_divorced = 1 vs 0 (within age $>$ 60)	Divorced split from not divorced	Among customers $>$ 60, not-divorced customers are more likely to subscribe (conditional on later splits).

Level 3	education_ternary	education_ternary = 1 vs 0 (within age \leq 60 & single)	Tertiary split from non-tertiary	Within younger single customers, tertiary education increases predicted probability.
Level 3	age	age \leq 31 (within age \leq 60 & single)	Younger split from older	Among single customers \leq 60, age \leq 31 forms a higher-conversion subgroup.
Level 3	job_retired	job_retired = 1 vs 0 (within age $>$ 60 & not divorced)	Retired split from non-retired	Within older not-divorced customers, being retired further increases predicted probability.

2.2 Targetable Customer Segments (Leaf Rules)

Using leaf-node rules (sample count, population share, and positive rate), I grouped customers into actionable segments.

Table 2.2. Summary of customer segments and conversion rates (selected leaf rules).

samples	proportion	pos_rate	rule
30	0.003	0.467	age $>$ 60 AND marital_divorced = 1
47	0.005	0.447	age $>$ 60 AND marital_divorced = 0 AND job_retired = 1
142	0.016	0.415	age $>$ 60 AND marital_divorced = 0 AND job_retired = 0

918	0.102	0.184	$\text{age} \leq 60 \text{ AND}$ $\text{marital_single} = 1$ $\text{AND } \text{age} \leq 31$
1637	0.181	0.126	$\text{age} \leq 60 \text{ AND}$ $\text{marital_single} = 1$ $\text{AND } \text{age} > 31$
1677	0.185	0.120	$\text{age} \leq 60 \text{ AND}$ $\text{marital_single} = 0$ AND $\text{education_tertiary} = 1$
4592	0.508	0.084	$\text{age} \leq 60 \text{ AND}$ $\text{marital_single} = 0$ AND $\text{education_tertiary} = 0$

Notes: samples = number of customers in the segment; proportion = segment share of the full dataset; pos_rate = observed subscription rate within the segment.

Segment interpretation

- High-value segment (high conversion, small size):
 - Customers with age > 60 show substantially higher subscription probability (pos_rate ~ 0.41–0.47). Although they represent about 2.4% of the sample, they are highly convertible and suitable for focused outreach (e.g., relationship-manager calls or personalised offers).
- Medium-value segments (moderate conversion, broad coverage):
 - Customers age ≤ 60 who are single (especially age ≤ 31), and customers age ≤ 60 who are non-single with tertiary education show moderate conversion and together cover a large share of customers. These segments are appropriate for scaled digital channels (SMS/email/app messaging) when budget allows.

3. Validating Drivers: Permutation Importance and SHAP

To validate the tree-based insights, I examined feature importance using grouped permutation importance (drop in ROC-AUC when shuffling features) and SHAP values (feature contributions to individual predictions).

3.1 Grouped Permutation Importance

Table 3.1. Grouped permutation importance (ΔAUC ; higher means more important).

Group	ΔAUC (mean)	\pm (std)	# dummy cols
Age	0.0784	0.0061	1
Education	0.0424	0.0059	4
Marital status	0.0268	0.0046	3
Job	0.0191	0.0042	12

- Age is the dominant predictor (largest ΔAUC).
- Education is consistently useful.
- Marital status contributes modestly and is most informative through interactions with age.
- Job importance is diffuse across many categories, yielding smaller per-feature contributions.

3.2 SHAP Analysis

Figure 3. SHAP summary plot (global feature effects).

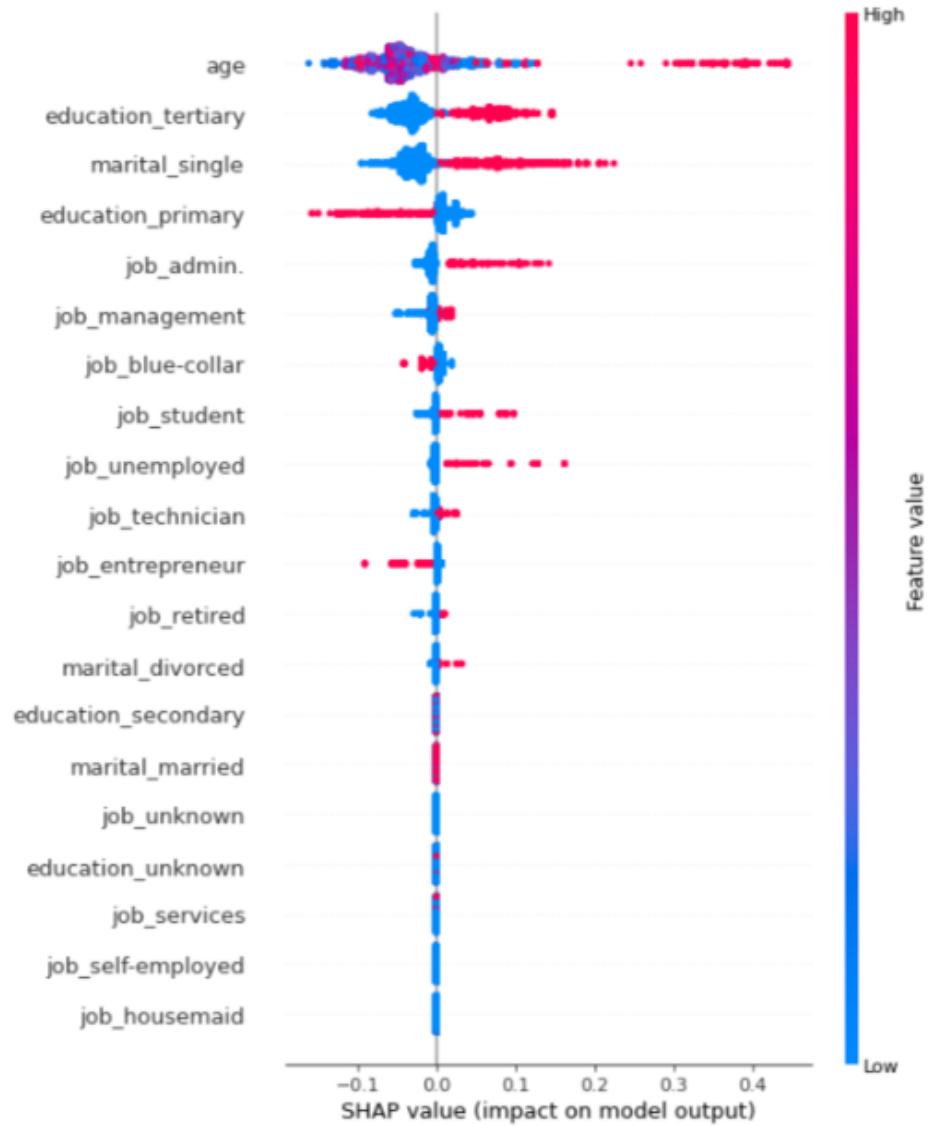
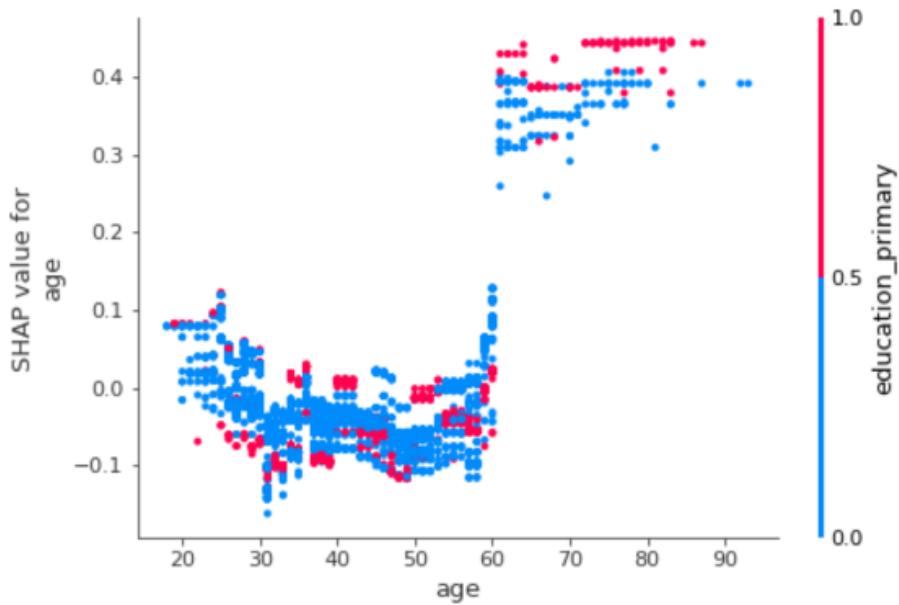


Figure 4. SHAP dependence plot for age (coloured by education_primary).



- Age has the strongest effect: SHAP contributions become clearly positive after ~60, while mid-age ranges (roughly 30–55) are mostly negative.
- Education shows a clear direction: tertiary education increases subscription likelihood, while primary education decreases it.
- Marital status: globally, being single is slightly negative on average; however, the tree indicates conditional uplift for younger single customers, highlighting interaction effects.
- Job effects are comparatively small, with some categories (e.g., admin/management/student) mildly positive and blue-collar slightly negative.

Conclusion

Across the decision tree, permutation importance, and SHAP analysis, age, education, and marital status emerge as the most informative demographic drivers. The analysis identifies a high-conversion senior segment ($\text{age} > 60$) and several medium-value segments that can support targeted marketing. However, performance metrics suggest that demographics alone are unlikely to deliver substantial predictive lift; incorporating financial and behavioural signals would be the next step for improved targeting.