Project Overview: Bank Subscription Prediction

This project aims to predict whether a customer will subscribe to a term deposit product based on historical banking campaign data. A classification model was built using a well-structured machine learning pipeline that integrates **feature engineering**, **data preprocessing**, **imbalanced learning techniques**, and **model optimization**.

1. Exploratory Data Analysis (EDA)

A. General Subscription Overview

- Only **11.7% of contacted clients subscribed**, highlighting a **very low baseline conversion rate**.
- **88.3% rejected the offer**, emphasizing the need for **better targeting and campaign design**.

B. Customer Profile Insights

Feature	Insight
Job	Students (2.1%) and retirees (5%) have the highest conversion rates. Blue-collar workers (21.5%) are the largest group but the least responsive.
Default Status	98.2% have no credit default — representing a **low-risk target base**.

Feature	Insight
Housing Loan	44% have no housing loan — indicating **higher disposable income**.
Previous Contact	Clients with a **previous successful campaign history** have higher conversion rates (23.1% vs. 9.2%).

C. Call Strategy & Duration

Metric	Impact on Subscription
Call Duration	**Strongest predictor** of subscription (correlation = **0.39**). Calls lasting **100–1000 seconds** yield the best outcomes.
Short Calls (<100s)	Very low conversion. Focus on **improving agent engagement tactics**.
Conversion Rate	Surges above **40%** for calls lasting **6+ minutes**.

Action: Train agents to maintain call durations in the 100s–1000s seconds range. Discard or revise scripts that end calls prematurely.

D. Campaign Fatigue

- Conversion rates peak with **0–3 campaign contacts** (~12–14%).

- Sharp **drop after 5+ contacts**, with negligible conversions after **10+ attempts**.

Action: Cap outreach attempts at 3. Beyond this, re-segment or retire the contact.

E. Monthly Performance & Trends

- **Peak Contact Months:** May–August (71% of all contacts).
- **Best Conversion Months:** March, June, October — despite **lower outreach volume**.
- **Worst Performing Month: May** — high volume, low results.

Action: Redistribute campaigns away from **May**. Focus on **seasonal timing** for better results.

F. Communication Channels

- **Cellular contacts dominate (65%)** and show **better conversion rates**.
- Customers contacted via "unknown" methods **almost never convert**.

Action: Prioritize cellular outreach. Remove or reassess other contact methods.

G. Education & Subscription Behavior

Education Level	**Conversion Trend**
Tertiary (Higher)	Most likely to subscribe — prefer detail-rich financial offerings.
Secondary	Largest group, but highest rejection rate.
Primary/Unknown	Least likely to subscribe.

Action: Tailor marketing content with **financial depth for tertiary group**. Use simplified messaging for lower education groups.

H. Other Noteworthy Patterns

- **Single Individuals** are more likely to subscribe → ideal for **youth-targeted campaigns**.
- **No Personal Loan** = higher openness to investment offers.
- **Returning Clients** (those previously contacted) subscribe **2.5x more** than first-timers.

I. Correlation Summary

Variable	Correlation with Subscription	Implication
Duration	+0.39	Best predictor; use to model lead quality or agent performance.
Previous Contacts	+0.093	Modest impact — could be modeled in campaign frequency logic.
Pdays (last contact)	+0.10	Slight positive — could help segment warmer leads.

Variable	Correlation with Subscription	Implication
Age, Balance, Day	Near 0	Not predictive — exclude from modeling decision-making.

2. Methods Used

Data Preprocessing:

- Removed unrealistic features (duration, day, month, contact, default) unavailable before prediction.
- Handled outliers in balance and pdays via capping or log-transformation.
- Replaced "unknown" in job, education, poutcome with np.nan for imputation with most frequent values.
- Applied StandardScaler to numeric features (campaign, engagement_score) and one-hot encoding to categorical features.

3.. Feature Engineering & Preprocessing

Implemented via a custom FeatureEngineer class and scikit-learn pipelines.

Feature Engineering Steps:

- was_contacted: Binary flag for if customer had been contacted before.
- engagement_score: Ratio of previous contacts to days since last contact.

- age_group: Binned into young, mid, senior, retired.
- Balance_Tier: Binned into negative, low, medium, high.

Preprocessing Pipeline:

- **Numerical Features**: Imputed with median, standardized.
- **Categorical Features**: unknown values replaced and imputed, then one-hot encoded.
- **Binary Features**: Transformed from yes/no to 1/0.
- **Resampling**: Applied SMOTEENN to balance classes by oversampling the minority and cleaning ambiguous points.

4. Modeling Approach

🔍 Model: RandomForestClassifier

- Handles non-linearity, robust to multicollinearity, interpretable through feature importances.
- Class weights were set to 'balanced' to improve sensitivity to the minority class.

Pipeline:

- Integrated preprocessing, resampling, and classification into one robust scikit-learn pipeline.

Hyperparameter Tuning:

- Conducted using GridSearchCV with cross-validation (cv=3).
- Parameters tuned:

- n_estimators: [100, 150]
- max_depth: [10, 15]
- min_samples_leaf: [3, 5]
- SMOTE k_neighbors: [3, 5]

5. Model Performance

Evaluated on a hold-out test set:

Metric	Value
F1 Score	~0.79
Recall	~0.81
Precision	~0.76

Interpretation:

The model performs well at **identifying customers likely to subscribe**, with good balance between precision and recall. This is critical in marketing as false positives (predicting subscription when it won't happen) are costly, but false negatives (missing a potential subscriber) are worse.

6. Model Deployment (Prediction Pipeline)

- The model and feature engineering logic are saved using joblib.

- In prediction.py, new customer data can be passed to the model for real-time prediction.
- It returns both the **predicted class** and **probability of subscription**.

Sample Input:

```
{
  'age': 40,
  'job': 'admin.',
  'marital': 'married',
  'education': 'secondary',
  'balance': 2000,
  'housing': 'yes',
  'loan': 'no',
  'campaign': 2,
  'pdays': 10,
  'previous': 2,
  'poutcome': 'success'
}
```

Output:

- **Prediction**: 1 (Subscribed)
- **Probability**: e.g., 0.82 (82% confidence)

7. Key Insights for Stakeholders

- **Engagement score** and **past campaign success** are strong predictors of subscription.
- **Customer financial health** (balance tier) positively influences outcomes.
- **Data quality matters**: handling unknowns and imputations was key.
- **Balanced approach** using SMOTEENN provided superior results over naive oversampling.
- **Model can be used** to optimize campaign targeting, reducing costs and improving conversions.

Limitations

- **Class Imbalance**: The dataset is inherently imbalanced, with significantly more "No" (non-subscribers) than "Yes" (subscribers) responses. Despite using resampling techniques like **SMOTEENN**, the model still exhibits bias towards the majority class.
- **Overfitting to 'No'**: The model appears to have learned the "No" class more effectively than the "Yes" class, resulting in lower precision or recall for identifying true subscribers. This may impact its usefulness in targeting potential customers who are likely to subscribe.