BANK TERM DATA ANALYSIS REPORT

SUMMARY

This report presents a comprehensive analysis of banking data to predict client subscription to term deposits. Using machine learning techniques on a dataset of 41,188 records, we developed predictive models with a focus on key banking features to achieve optimal performance for production deployment.

1. DATA OVERVIEW

Dataset Characteristics

- **Total Records**: 41,188 client interactions
- **Original Features**: 20 input variables plus target variable
- **Target Variable**: Binary (yes/no) term deposit subscription converted to (0/1)
- **Class Distribution**: Highly imbalanced dataset (approximately 11.3% positive class)

Data Quality Assessment

- **Missing Values**: None detected
- **Duplicate Records**: 12 duplicates found
- **Data Types**: Mixed (numerical and categorical)
- **Unknown Values**: Present in categorical features (job, marital, education, poutcome)

2. EXPLORATORY DATA ANALYSIS (EDA)

Target Variable Distribution

The analysis revealed a significant class imbalance in the target variable, with the majority of clients not subscribing to term deposits.

Categorical Features Analysis

Comprehensive analysis was conducted on all categorical variables including:

- Job categories with their distribution patterns
- Marital status frequencies
- Education levels
- Previous campaign outcomes (poutcome)

Correlation Analysis

A correlation matrix was generated for numerical features to identify relationships and potential multicollinearity issues among the variables.

3. FEATURE ENGINEERING

Feature Selection Strategy

Instead of creating new features, the analysis focused on selecting the most relevant features for banking marketing campaigns:

python

```
# Key features for banking marketing campaigns
key_features = [
    'age', 'job', 'marital', 'education',
    'campaign', 'pdays', 'previous', 'poutcome',
    'emp.var.rate', 'cons.price.idx', 'cons.conf.idx',
'euribor3m'
```

```
]

# Select only key features + target
df_focused = df[key_features + ['y']].copy()
```

Data Preprocessing Steps

1. Unknown Value Handling

python

```
# Handle unknown values in categorical columns
categorical_cols = ['job', 'marital', 'education', 'poutcome']
for col in categorical_cols:
    # Replace 'unknown' with mode
    mode_value = df_focused[df_focused[col] !=
'unknown'][col].mode()
    if len(mode_value) > 0:
        df_focused[col] = df_focused[col].replace('unknown',
mode_value[0])
```

2. Categorical Encoding

python

```
# Prepare data for modeling
# Encode categorical variables using one-hot encoding
df_encoded = pd.get_dummies(df_focused, drop_first=True)

# Define features and target
X = df_encoded.drop('y', axis=1)
y = df_encoded['y']
```

3. Class Balancing

python

```
# Handle class imbalance with SMOTE
from imblearn.over_sampling import SMOTE

smote = SMOTE(random_state=42)
X_train_balanced, y_train_balanced = smote.fit_resample(X_train,
y_train)
```

4. MODEL DEVELOPMENT

Models Implemented

Three different machine learning models were trained and evaluated:

1. **Random Forest Classifier**
    - n_estimators=100
    - class_weight='balanced'
    - random_state=42
2. **XGBoost Classifier**
    - n_estimators=100
    - scale_pos_weight calculated based on class distribution
    - eval_metric='logloss'
3. **Logistic Regression**
    - class_weight='balanced'
    - max_iter=1000

Training Strategy

- **Data Split**: 80% training, 20% testing with stratified sampling
- **Class Imbalance Handling**: SMOTE oversampling applied to training data
- **Evaluation Metrics**: Classification report, ROC AUC score, confusion matrix

6. MODEL PERFORMANCE AND RESULTS

Model Evaluation Process

Each model was evaluated using a comprehensive evaluation function that provided:

- Classification reports with precision, recall, and F1-scores
- ROC AUC scores for model comparison
- Confusion matrices for detailed performance analysis
- Feature importance rankings

Best Model Selection

The best performing model was selected based on ROC AUC score, with all model results stored for comparison.

8. FEATURE IMPORTANCE ANALYSIS

Top Feature Insights

The analysis identified the most important features for predicting term deposit subscriptions. The top 10 features from the best performing model were ranked and displayed, providing insights into which variables have the strongest predictive power.

Key categories of important features typically include:

- **Economic indicators**: Macroeconomic factors affecting client decisions
- **Previous campaign data**: Historical interaction patterns
- **Demographic factors**: Age, Marital status and other client characteristics

9. BUSINESS INSIGHTS AND RECOMMENDATIONS

Key Findings

Based on the analysis, feature importance analysis and model performance:

1. **Call Duration Significance**: Longer call durations indicate higher client interest and engagement
2. **Economic Indicators Impact**: Macroeconomic conditions significantly influence subscription decisions
3. **Previous Campaign Outcomes**: Historical interaction data provides valuable predictive power
4. **Contact Frequency Optimization**: Campaign frequency should be balanced to avoid client fatigue

Strategic Recommendations

1. **Focus on Call Quality**: Prioritize meaningful, engaging conversations over quantity
2. **Economic Timing**: Consider macroeconomic indicators when planning campaign timing
3. **Leverage Historical Data**: Use previous campaign outcomes to inform targeting strategies
4. **Optimize Contact Strategy**: Balance contact frequency to maximize effectiveness

10. CONCLUSION

The analysis led to the creation of a strong predictive model for term deposit subscriptions, built on thoughtful feature engineering and model evaluation. By prioritizing key banking features, addressing class imbalances effectively, and thoroughly comparing model performance, the solution is both reliable and deployable.

This model offers valuable insights for marketing teams, enabling them to refine campaign strategies, allocate resources more efficiently, and boost conversion rates.