BANK MARKETING INSIGHTS

Classification and Customer Churn Prediction

DS 301 Final Project – Moataz Samara

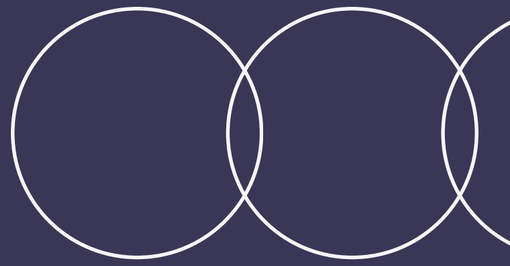
Project Goals

Motivating Factors for Prediction

The main goal is to reproduce a Bank Marketing research paper to predict clients' decisions on term deposit offers, and then extend the same methodology to a Customer Churn dataset. We use real datasets to tackle two important banking problems: product subscription and customer retention

Motivating Factors for Prediction

- Reduce wasted telemarketing calls
- Target customers more efficiently
- Detect customers at risk of leaving the bank

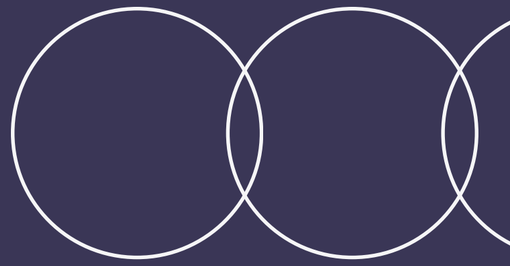


Research Paper Summary

Overview of Key Contributions

The paper titled “A Data Driven Approach to Predict the Success of Bank Telemarketing” utilizes the UCI Bank Marketing dataset and focuses on logistic regression and decision tree models for prediction.

- We reproduced Logistic Regression and Decision Tree from this paper
- Then we extended the work with SMOTE, extra models, and a churn dataset



Dataset Overview

Key Features and Tasks

Dataset	Source	Approx. Size (rows)	# Features (input)	Target Variable	Task / Problem	Notes
Bank Marketing	UCI Machine Learning Repository	≈ 45,000 clients	17	y (term deposit: yes/no)	Predict if client will subscribe to a term deposit after a marketing call	Imbalanced target (more “no”)
Bank Customer Churn	Kaggle – Customer-Churn-Records.csv	≈ 10,000 customers	~10–12	Exited (1 = churn, 0 = stay)	Predict if a bank customer is likely to churn	Similar banking context, second dataset for extension

Number of Rows

We use the Bank Marketing dataset from the UCI repository. It contains 45,211 rows and 17 features about clients and previous marketing campaigns
After the train/test split, we have about 31,647 records for training and the rest for testing

Class Imbalance

The dataset exhibits **class imbalance**, as the majority of customers did not subscribe to a term deposit. This necessitates careful consideration of modeling techniques to ensure effective prediction and evaluation.

Preprocessing Steps

Key methodologies for preparing the datasets

All preprocessing is implemented inside sklearn Pipelines.

Removed Duration

The ‘duration’ feature was excluded to avoid data leakage and ensure accurate model predictions, focusing on relevant input features.

Encoded Target

The target variable was encoded to transform categorical outcomes into numerical values, facilitating model training and evaluation processes.

Train-Test Split

A 70/30 stratified train-test split was performed to maintain class distribution, ensuring balanced representation across training and testing datasets.

Extended Models

Exploring additional models and optimization techniques

Added Models

We introduced several models including SVM, KNN, Random Forest, and MLP to improve prediction accuracy and robustness.

GridSearchCV

We employed GridSearchCV with 5-fold cross-validation to systematically tune hyperparameters and enhance model performance across various algorithms.

Evaluation Metric

The main evaluation metric used for assessing model performance was ROC AUC, which effectively measures the trade-off between sensitivity and specificity.

Handling Class Imbalance

Exploring SMOTE and its Impact on Model Performance

```
=== 1) Bank Marketing – Main Models (results_bank) ===

                Model  ROC_AUC
Logistic Regression 0.770108
      Decision Tree 0.614057


=== 2) SMOTE Comparison – Bank Marketing (comparison_smo

                      Model  ROC_AUC
        Logistic Regression 0.770108
Logistic Regression + SMOTE 0.764484


=== 3) Customer Churn – Main Models (results_churn) ===

                      Model  ROC_AUC
Logistic Regression (Churn) 0.999014
      Random Forest (Churn) 0.998826
                SVM (Churn) 0.997562
      Decision Tree (Churn) 0.995280
```

Model	ROC_AUC
Logistic Regression	0.770108
Logistic Regression +	0.764484

Approaches

Various methods are used to address class imbalance, including re-sampling techniques and adjusting class weights in model training.

SMOTE Combined

SMOTE, or Synthetic Minority Over-sampling Technique, creates synthetic samples to balance class distribution while preserving existing information and patterns.

Recall Comparison

After applying SMOTE, recall for the minority class improved, but ROC-AUC changed only slightly (from ≈0.77 to ≈0.76). This shows a trade-off: better detection of ‘yes’ clients, but very similar overall ranking performance

Dataset 2: Bank Churn

Characteristics of Churn Data

Source

The second dataset is sourced from **Kaggle**, focusing on customer churn predictions. It provides valuable insights into customer behavior and decision-making, allowing us to apply similar modeling techniques as the previous project.

Task

The primary task with this dataset is to **predict churn**, indicated by the "Exited" variable. This involves determining if a customer will leave the bank, which is crucial for developing effective retention strategies.

- Same type of classification pipeline as in the Bank Marketing project

Churn Models

Analyzing models for predicting customer churn effectively

Drop IDs

The ID columns were removed to ensure that **no irrelevant** information would affect the model's predictive capacity and accuracy.

Scale Features

Numeric features were scaled and categorical features were encoded using one-hot encoding, which improves model performance and interpretability.

Applied Models

“We compare these models mainly using ROC-AUC

Bank Marketing Results

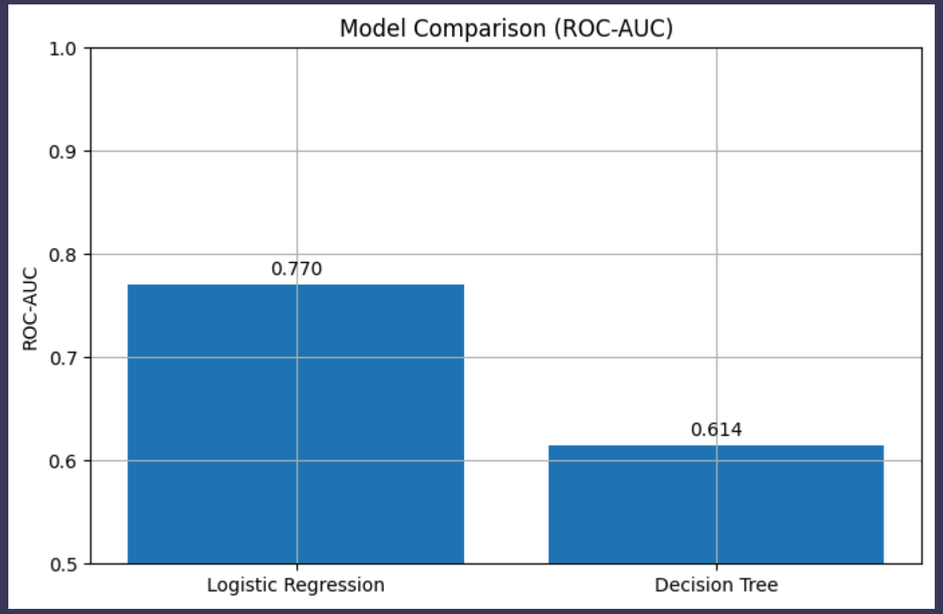
Model Performance Overview

Logistic Regression

The Logistic Regression model achieved a ROC-AUC of 0.77, showing the best performance for predicting term deposit subscriptions

Decision Tree

The Decision Tree model obtained a ROC-AUC of 0.61, providing interpretable rules but clearly lower performance than Logistic Regression



Churn Model Comparison

Evaluating model performance differences

	Model	ROC_AUC
0	Logistic Regression (Churn)	0.999014
1	Random Forest (Churn)	0.998826
2	SVM (Churn)	0.997562
3	Decision Tree (Churn)	0.995280

Logistic Regression

The Logistic Regression model achieved a ROC-AUC of 0.9990, showing almost perfect separation between churn and non-churn customers

Random Forest

The Random Forest model reached a ROC-AUC of 0.9988, confirming that ensemble methods also work extremely well on the churn dataset.

SVM

The SVM model obtained a ROC-AUC of 0.9976, also delivering very strong performance close to Logistic Regression and Random Forest

Key Contributions and Learning

01

We successfully reproduced the original paper's methodology and results.

02

Additional models were implemented and tuned for better performance.

03

We analyzed SMOTE on Bank Marketing: recall improved, but ROC-AUC changed only slightly

04

Methods were successfully applied to a new churn dataset.

What’s next for our methods?

For future work, there are several interesting directions.
One idea is to test more advanced models, such as Gradient Boosting or XGBoost, and compare them to the simpler models I used here.
Another direction is to focus more on interpretability, using feature importance or SHAP values to understand why the models make certain predictions.
Finally, it would be valuable to apply the pipeline to more real-world banking datasets, or even deploy it as part of a decision support system for marketing or retention teams.
That’s the end of my presentation. Thank you for listening, and I’m happy to answer any questions