

Project Writeup

Classification Project Write-up

Predicting the Bank Telemarketing Success

Abstract

The goal of this project was to use classification models to predict the bank telemarketing success. I worked with data provided by [UCI Machine Learning Repository](#)(source — [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014). I built several classifiers to test the data and arrived at a tuned Random Forest model with the best precision score on positive class. I also suggest using the random forest model with probability threshold adjustment so that depending on the business sources bank have (which decides how many clients bank can reach to or how many phone calls bank can do in a given period (week/month/year), the model can have the exact amount of the number of clients bank will reach out to(predicted positives), and number of success(true positives) suggested by the model.

Design

The goal of the project is to classify whether a client will subscribe the bank term deposit through telemarketing campaign. Yes is the positive class. A more specific goal is to improve the precision score on positive class and have a relative large number of true positives from the predicted positives.

Data

“The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.” The original data consists of 45211 rows, 20 features and 1 target. I narrowed it down to 15 relevant features for the project. Target is yes/no subscribe or 1/0 as mentioned above.

Features contains bank clients' data including:

1 - age: numeric, 18 to 95.

2 - job : type of job (categorical:

"admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services")

3 - marital : marital status (categorical: "married", "divorced", "single"; note: "divorced" means divorced or widowed)

4 - education (categorical: "unknown", "secondary", "primary", "tertiary")

5 - default: has credit in default? (binary: "yes", "no")

6 - balance: average yearly balance, in euros (numeric)

7 - housing: has housing loan? (binary: "yes", "no")

8 - loan: has personal loan? (binary: "yes", "no")

And campaign data and other attributes

9 - contact: contact communication type (categorical: "cellular", "telephone", "unknown")

10 - day: last contact day of the month

11 - month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)


15 - poutcome: outcome of the previous marketing campaign (categorical: "failure", "nonexistent", "success")

Data date ranges from May 2008 to November 2010.

Algorithms

I label encoded categorical features and fit several classifiers as following:

Modeling & Performance Metric Report

Phase 1: Model Testing				
Model	Precision	Recall	F-1	Accuracy
1. K-Nearest Neighbor Baseline	0.43	0.12	0.18	0.88
2. K-Nearest Neighbor Optimized with Grid Search	0.48	0.04	0.07	0.88
3. Logistic Regression Baseline	0.50	0.00	0.00	0.88
4. Logistic Regression Regularized	0.48	0.04	0.07	0.88
5. Random Forest Baseline	0.69	0.21	0.32	0.89
6. Random Forest Optimized with Random Search	0.71	0.19	0.30	0.89
Phase 2: Handle Class Imbalance				
Model	Precision	Recall	F-1	Accuracy
7. Random Forest with Sampling method	0.55	0.29	0.38	0.89
8. Random Forest with Adjusted Class Weight	0.69	0.19	0.30	0.89
9. Random Forest with Probability Threshold Adjustment 	Adjustable	Adjustable	Adjustable	Adjustable

Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for visualization

Communication

See presentation pdf.