Using Classifiers to Identify Factors Contributing to Marketing Campaign Success

Overview

The performance of four classifiers (k-nearest neighbors, logistic regression, decision trees, and support vector machines) was compared to a base model for predicting customers subscribing a term deposit. Factors contributing to customers saying 'yes' to marketing campaigns were identified, including customer characteristics, campaign timing, and frequency of contact. Successful campaigns targeted highly educated customers who had said yes in a previous campaign and spent the longest time with agents on the phone. They were contacted mid-week in the spring or fall at times when the Euribor rate and Consumer Price Index were both high.

Business Understanding

Objective

The business objective of this project is to identify the factors that most influenced the results of marketing campaigns conducted by a Portuguese bank. Identifying the factors most related to success will allow better targeting of prospective clients and management of firm resources in marketing efforts, increasing efficiency and saving costs.

Data Mining Goals

The research and data mining goals are to identify the most important predictors of success and to compare the performance of the classifiers (k-nearest neighbors, logistic regression, decision trees, and support vector machines) in predicting customers subscribing a term deposit.

Resources

The dataset results from 17 marketing campaigns that occurred between 2008 and 2010. It comes from the UCI Machine Learning repository link. This analysis was conducted in python using pandas and sklearn tools on a Jupyter notebook, link. If follows the CRISP-DM project format.

Data Understanding

Data Description

The dataset contains 41,188 cases representing customers with 20 numerical and categorical input features and 1 binary target variable representing ‘yes’ or ‘no’. There are 10 client attributes, 5 relating to the marketing campaigns, and 5 attibutes representing social and economic context factors. More details may be found at the repository link.

Data Cleaning

Many of the variables were categoricals which needed to be numerically encoded prior to modeling, including the target feature "y". There were no duplicates or outliers. Several numeric features had skewed or discontinuous distributions which required scaling.

The target variable is categorical and will be converted to numeric. The dataset authors warned that the feature 'pdays' is highly correlated with 'no' and should be dropped, so we dropped it. The cleaned dataset had 19 features.

While no null values were present in the dataset, some cases had missing data represented by the category "unknown". The category value counts for each feature showed that the percentage of "unknown" was low for most of the features. The exception was the feature "default", which was unknown for 16% of the data. Since we have a large dataset we went ahead and dropped all the cases with unknown/missing data, rather than using imputation.

After dropping all cases with missing values, the cleaned dataset contained 30,488 cases. Of these, 3,859 were ‘yes’, a success rate of 12.7%.

Modeling

Model Test Design

1. Preparation. Create training and validation sets
2. Preliminary models. Compare classifiers to a baseline model using default settings
3. Model Improvement. Feature selection and engineering and parameter hypertuning using GridSearch

Preparation

The data was split into target and independent features, and then into training and test sets, with a ratio of 30% for testing.

A baseline model was calculated using the Dummy Classifier tool, and evaluated using an accuracy score. The baseline score was .87.

A preprocessor was created to transform the independent features. Numerical features were regularized using StandardScalar and dummy variables were created from the categorical variables using OneHotEncoder.

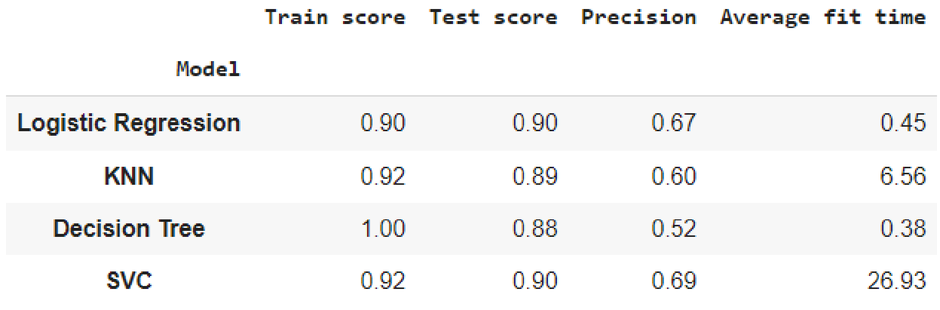
Preliminary Models

Four classification methods (k-nearest neighbors, logistic regression, decision trees, and support vector machines) were tested using default settings in a Pipeline with the preprocessor. Fit times and evaluation metrics were recorded and entered into a summary table, shown in the table below.

The accuracy score for the initial logistic regression model was .90. The precision score was .92 for 'no' and .68 for 'yes'. Since we care more about accurate prediction of 'yes' outcomes, we will try other classifiers to see if they can improve accuracy.

First, the K Nearest Neighbors classification method was tried. The KNN accuracy score was .89 and precision was .61, so it was not a better classifier. Next, a Decision Tree method was tried. The DT accuracy score was .88 and precision score was .51, meaning that it did a much worse job. The initial DT had a depth of 27 levels. Finally, a Support Vector Classifier method was tried. The SVC accuracy score was .90 and the precision score was .69. That means it outperformed the other models, just barely. However it took a very long time to estimate.

Preliminary Results



	Train score	Test score	Precision	Average fit time
Model				
Logistic Regression	0.90	0.90	0.67	0.45
KNN	0.92	0.89	0.60	6.56
Decision Tree	1.00	0.88	0.52	0.38
SVC	0.92	0.90	0.69	26.93

Model Improvement

We took the following steps to improve the models, optimizing for precision:

- Logistic Regression - Use SelectFromModel tool to look at model coefficients and see if any low value ones that could be dropped

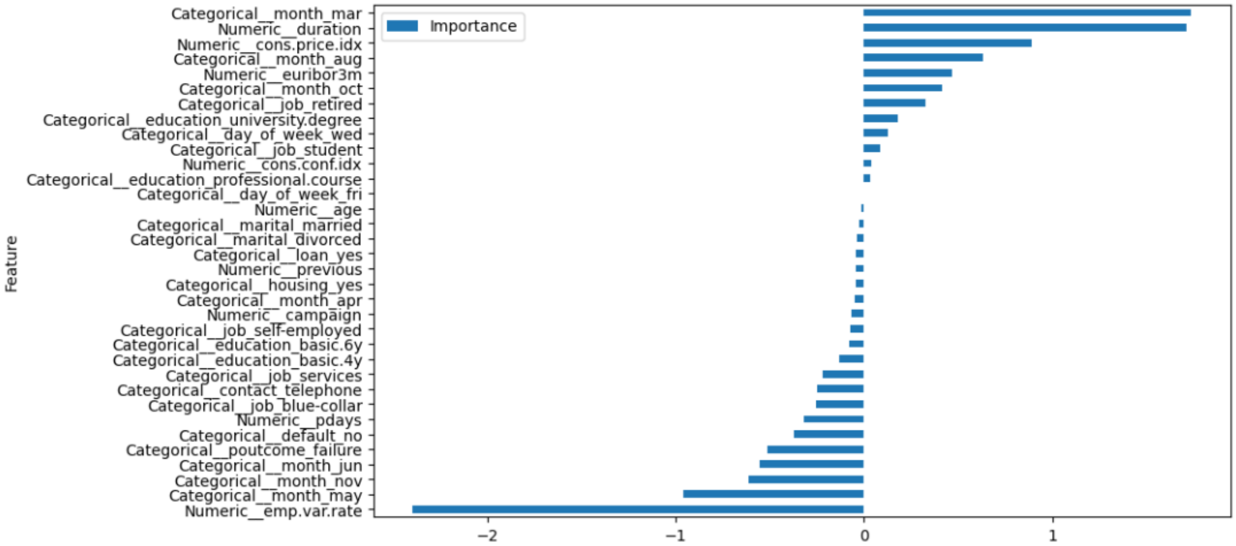
- K Nearest Neighbors model - grid search of k-values
- Decision Tree - grid search of max tree depths
- Support Vector Classification - grid search gamma values and kernels

Feature Selection

Since this dataset has 54 predictor features, which can just add noise and slow down processing, it was important to identify the most and least important ones. Reducing noise and degrees of freedom should help improve predictive power of the models. Using the SelectFromModel tool, we used logistic regression with an L1 (Lasso) penalty and a C value of .01 as a threshold coefficient value. Since the 'yes' class of interest was only 12% of the data, the parameter for class weight was set to 'balanced' to automatically adjust weights inversely proportional to class frequencies.

As a result of these constraints, the top 36 features were selected by the tool. They are shown below in order of importance, by coefficient value. A full list with coefficient values may be found in the Appendix. What the most important features tell us about the marketing campaigns is discussed in the Results section.

Feature Selection Results



Feature Engineering

One feature, 'education', had many categories but only a few made it to the most important list. The one with the highest predictive power of success was converted into a dummy variable called university_degree.

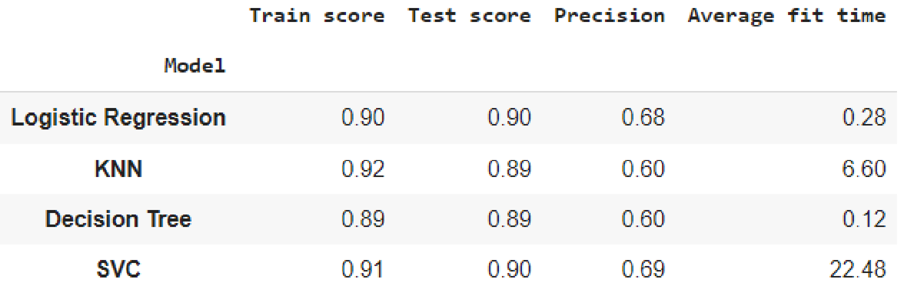
Features that were not selected as having high predictive power were dropped, including: age, campaign, contact, loan, housing, marital, emp.var.rate, and education.

The resulting revised dataset had 12 features. The pre-processor step was revised to account for fewer features in the input arrays.

Parameter Hypertuning

The GridSearch tool was used to tune important parameters of each modeling method. After optimal parameters were identified, each model was re-run using those values. Fit times and evaluation metrics were recorded and entered into a summary table, shown in the table below.

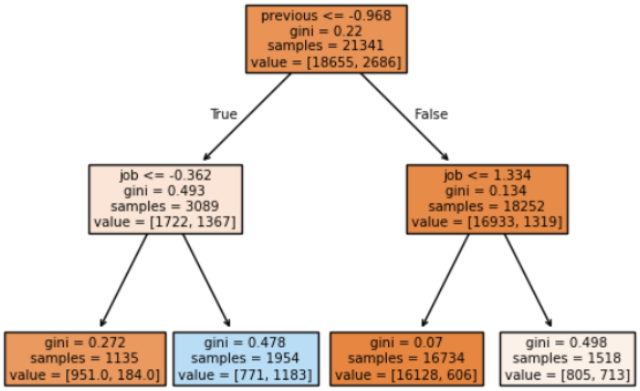
Results from Improved Models



	Train score	Test score	Precision	Average fit time
Model				
Logistic Regression	0.90	0.90	0.68	0.28
KNN	0.92	0.89	0.60	6.60
Decision Tree	0.89	0.89	0.60	0.12
SVC	0.91	0.90	0.69	22.48

The grid search for the optimal k in the K Nearest Neighbors model was found to be 19. The revised KNN model had an accuracy of .89 and a precision of .60. The grid search for the Decision Tree model identified three optimal parameter values: a maximum depth of 2, minimum impurity decrease of .01, and minimum samples split of .1. The grid search for the Support Vector Classification model identified two optimal parameter values: an rbf kernel and gamma of .1. The revised SVC model had an accuracy of .91 and precision of .69. The revised decision tree model had an accuracy score of .89 and a precision score of .60. The top two levels were 'previous' and 'job', meaning these two factors had the highest level of predictive power.

Decision Tree Results



Evaluation

This analysis has achieved the primary client objective of identifying the most influential factors on marketing campaign success, in this dataset. This will allow better targeting of prospective clients and management of firm resources in marketing efforts, increasing efficiency and saving costs, as suggested in the recommendations below.

Findings

The efforts to improve the models did not yield much predictive benefit, in terms of improving accurate prediction of success. However there are a number of interesting findings about which customers are most likely to say 'yes' and which are most likely to say 'no', allowing for better targeting of resources, as shown in the summary table below. We can also observe the most effective campaign characteristics and timing.

Most influential factors towards success

Factor	**Likelihood of 'yes' increases with**	**Likelihood of 'no' increases with**	**Important but low impact**
Campaign	Previous campaign outcome success	Previous campaign outcome failure or nonexistent	Number of contacts previous campaign
	Duration of contact	Number of days since contact from a previous campaign	Number of contacts this campaign
		Contact by telephone	
Customer	University degree	Basic 4-year education	Basic 6-year or 9-year education, or professional course
	Retired or student job status	Blue collar or services job status	Self-employed job status
		No credit in default	Age or Marital status
			Has a personal or housing loan
Economy	Euribor rate	Employees Variation Rate	Consumer Confidence

			Index
	Consumer Price Index		
Time of year	March, August, October	May, June, November	April
	Wednesday		Friday

Customers most likely to say ‘yes’ had the highest level of education, a University degree, were students working towards one, and may have already retired. They had said yes in a previous campaign and spent the longest time with agents on the phone. They were contacted mid-week in the spring or fall at times when the Euribor rate and Consumer Price Index were both high.

Customers most likely to say ‘no’ had the lowest level of education, basic 4-year, and worked in blue collar or service jobs. They may have said no in a previous campaign or have not been contacted before. They were contacted in spring or fall by telephone during times when the Employee Variation Rate was high.

No other factors significantly impacted the likelihood of success, including other levels of education or types of jobs, age, marital status, or whether the customer held a housing or personal loan. The number of contacts by any campaign did not make a difference, nor did a high level of the Consumer Confidence Index.

Recommendations

Based on these findings, we recommend the following:

- Since previous campaigns were one of the most important predictors of success or failure, subsequent campaigns should
  - Cultivate deeper loyalty relationships with customers who say ‘yes’
  - Remove customers who said ‘no’ from further contact
- Since the number of contacts during each campaign did not significantly increase likelihood of success, and each contact has a cost, the number of contacts should be reduced, and
  - Reduce contacts by telephone
- Strategies targeting University students and retirees will be more successful
- Campaigns should be launched when the Euribor rate and Consumer Price Index are both high, and not launched when the Employees Variation Rate is high.
- The best months to conduct the campaign are March, Augst and October, and the best day to call a customer is Wednesday.

Next Steps

We recommend further data collection and modeling to improve predictions, due to the limited results of this effort. In particular, we need better data about the customer contact itself and specific messaging, since it appears that the agent had a significant influence on the outcome of a call.

Appendix

Ranked Table of Factors, Strength and Direction of Influence

Feature	Importance	Meaning
Categorical__month_mar	1.74	Month last contacted
Numeric__duration	1.71	Duration of last contact, in seconds
Categorical__poutcome_success	1.01	Outcome of previous marketing campaign
Numeric__cons.price.idx	0.89	Consumer price index (monthly)
Categorical__month_aug	0.64	Month last contacted
Numeric__euribor3m	0.47	Euribor 3 month rate (daily)
Categorical__month_oct	0.40	Month last contacted
Categorical__job_retired	0.32	Type of job
Categorical__education_university.degree	0.18	Highest level of education
Categorical__day_of_week_wed	0.12	Day of week last contacted
Categorical__job_student	0.11	Type of job
Numeric__previous	0.04	Number of contacts in previous campaign campaign
Numeric__cons.conf.idx	0.04	Consumer confidence index (monthly)
Categorical__education_professional.course	0.03	Highest level of education
Categorical__education_basic.9y	0.00	Highest level of education
Categorical__day_of_week_fri	-0.01	Day of week last contacted
Numeric__age	-0.02	Age
Categorical__marital_married	-0.03	Marital status
Categorical__loan_yes	-0.03	Has a personal loan?
Categorical__marital_divorced	-0.04	Marital status
Categorical__housing_yes	-0.05	Has a housing loan?
Categorical__month_apr	-0.06	Month last contacted
Numeric__campaign	-0.07	Number of contacts during this campaign
Categorical__education_basic.6y	-0.07	Highest level of education
Categorical__job_self-employed	-0.07	Type of job
Categorical__education_basic.4y	-0.13	Highest level of education
Categorical__job_services	-0.23	Type of job
Categorical__default_no	-0.24	Has credit in default?
Categorical__contact_telephone	-0.25	Contact type
Categorical__job_blue-collar	-0.26	Type of job
Categorical__poutcome_nonexistent	-0.27	Outcome of previous marketing campaign
Categorical__month_jun	-0.56	Month last contacted
Categorical__month_nov	-0.61	Month last contacted
Categorical__poutcome_failure	-0.90	Outcome of previous marketing campaign
Categorical__month_may	-0.96	Month last contacted
Numeric__emp.var.rate	-2.41	Number of employees (quarterly)