Predicting the success of marketing campaign

Kumararatnam Kabilesh

Introduction

Marketing campaigns constitute a typical strategy to enhance business. Companies use direct marketing when targeting segments of customers by contacting them to meet a specific goal. The banking sector is one of those industries that exercise direct marketing campaigns, in this case phone calls. Sometimes even though many marketing calls are made, only some of them may result in success. Therefore, predicting if the calls made result in success can help the marketing team to arrive at a conclusion of success of the campaign and adjust the number of calls accordingly. Also, the marketing team can change attributes or parameters based on past results to improve the success rate in future.

The data used in this study 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 ('yes') or not ('no') subscribed. The classification goal is to predict if the client will subscribe a term deposit (target variable y).

Data

As mentioned above, the dataset consists of direct marketing campaigns data of a banking institution. The dataset was picked from UCI Machine Learning Repository. There were four variants of the datasets.

1. bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]
2. bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
3. bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
4. bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).

This study was done using the ‘bank-additional-full.csv’ dataset which consists of 41188 data points with 20 independent variables out of which 10 are numeric features and 10 are categorical features. The list of features available to us are given below:

Input variables:

Attributes about bank client data

age, job, marital, education, default, housing, loan,

Attributes related with the last contact of the current campaign

Contract, month, day_of_week, duration, campaign, pdays, previous, poutcome

Social and economic context attributes

emp.var.rate, cons.price.idx, cons,.conf.idx, Euribor3m, nr.employed

Output variable (desired target):

y - has the client subscribed a term deposit? (binary: 'yes', 'no')

Methodology

This is a binary classification problem. Our two classes are “yes” denoting that the customer subscribed to a term deposit, and “no” denoting that the customer did not subscribe.

Exploratory data analysis and data preprocessing

Shape of data - (41188, 21)

Attribute	No of ‘unknown’	Data Type
age	int64	0
job	object	330
marital	object	80
education	object	1731
default	object	8597
housing	object	990
loan	object	990
contact	object	0
month	object	0

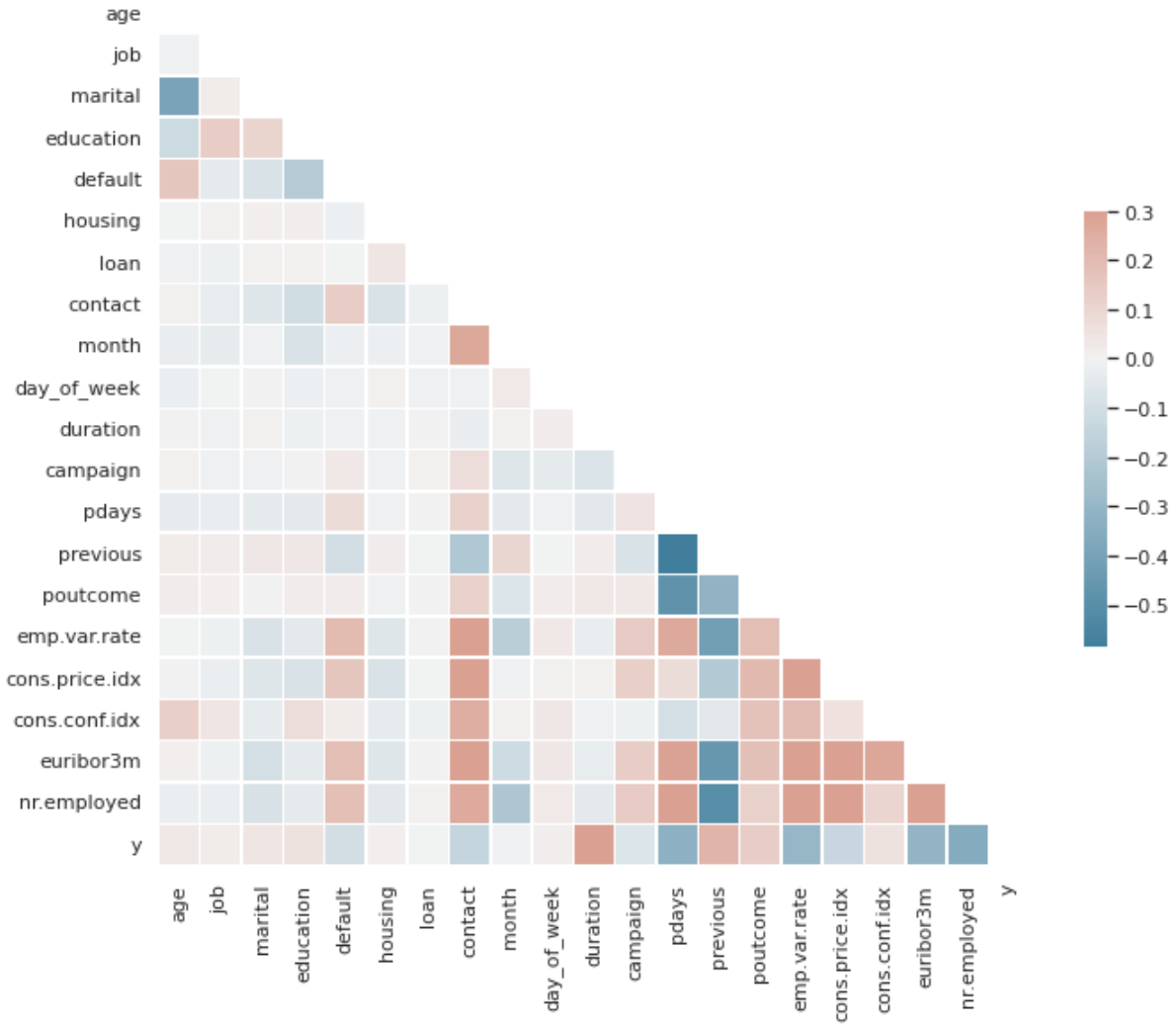
day_of_week	object	0
duration	int64	0
campaign	int64	0
pdays	int64	0
previous	int64	0
poutcome	object	0
emp.var.rate	float64	0
cons.price.idx	float64	0
cons.conf.idx	float64	0
euribor3m	float64	0
nr.employed	float64	0

The missing values had value as ‘unknown’ instead of ‘Nan’ or ‘null’. The number of missing values in each attribute can be seen in the above table.

Records with values as ‘unknown’ from attributes marital, job and housing were dropped from the data. This removed the rows with missing values for all attributes except ‘education’ (3.9% of records) and ‘default’ (20 % of rows). For both of these attributes, ‘unknown’ was used as an attribute value.

The dataset contains attributes with various data types such as int, float and object. Categorical attributes that are ordinal were encode with OrdinalEncoder and categorical attributes that are not ordinal were encoded with LabelEncoder.

Correlation between attributes



It is evident that some attributes have no correlation to the target variable. Therefore, those attributes can be omitted from the attributes for training the model.

Data was split in to train (70%) and test (30%) data. The following classification algorithms were trained for predicting the success of the marketing campaign.

```
from sklearn.ensemble import RandomForestClassifier
RandomForestClassifier(n_estimators=20, random_state=42)
```

```
from sklearn.linear_model import LogisticRegression
LogisticRegression(random_state=42, max_iter=250)
```

```
from sklearn.naive_bayes import GaussianNB
GaussianNB()
```

```
from sklearn.svm import SVC
SVC(gamma='auto')
```

```
import xgboost as xgb
XGBClassifier(objective ='binary:logistic', colsample_bytree = 0.3, learning_rate = 0.1,
              max_depth = 5, alpha = 10, n_estimators = 10)
```

The hyperparameters were tuned for the best model and, feature selection was applied.

```
from sklearn.feature_selection import RFE
RFE(estimator=model, n_features_to_select = 10, step = 4)
```

Results

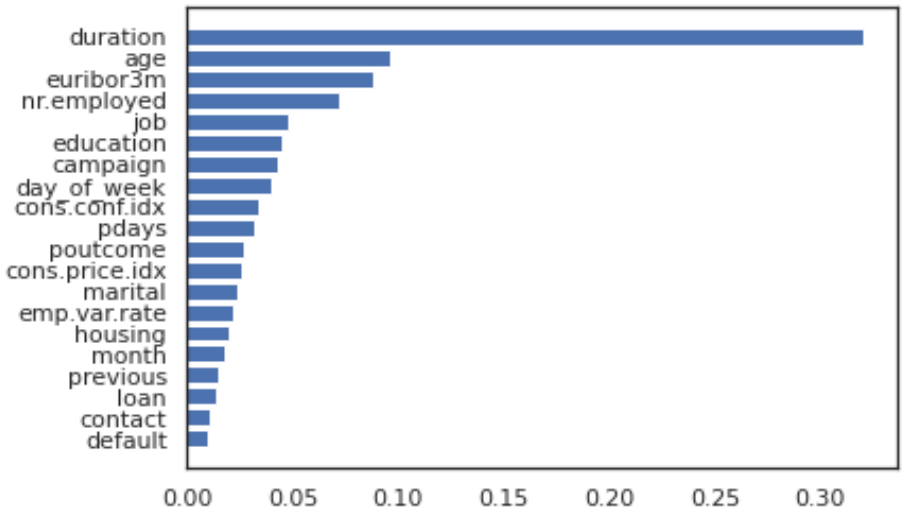
The accuracy score, precision, recall and f1-score were calculated for these predictions. The results were as following.

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.9077	0.7881	0.6853	0.6853
Naïve Bayes	0.8493	0.6697	0.7455	0.6951
Support Vector Machine	0.8870	0.6941	0.5058	0.4830
Random Forest	0.9136	0.7960	0.7326	0.7590
Random Forest with feature selection	0.9136	0.7921	0.7456	0.7660
XgBoost	0.9049	0.8380	0.6097	0.6503

Discussion

As it can be seen from the above table, ensemble based models and xgBoost perform the best. Random Forest with feature selection gave better performance than the one without feature selection.

Feature importance



Conclusion

The obtained results are credible for the banking domain and provide valuable knowledge for the telemarketing campaign manager. In our observations we concluded a classification model and applied identified the features or attributes which affect the outcome the most. The result of this is that the company could give valuable time and importance to a customer base which is largely going to affect their business. Such analysis could lead to huge savings in time and money for the relevant institution. Therefore, embracing machine learning can bring valuable insights and improve the efficiency of the business.