BANK MARKETING EFFICTIVENESS PREDICTION

K. YOGESH REDDY,

D. PRUTHVI RAJ

DATA SCIENCE TRAINEES,

ALMABETTER, BANGLORE.

ABSTRACT:

This dataset is provided by Portuguese banking institution. This deals about the direct phone call marketing to the existing customers, which aims to promote term deposit. The classification goal is to predict if the client will subscribe the term deposit(y).

Here we will perform different EDA techniques and with different classification algorithms to predict the output feature y(yes/no).

PROBLEM STATEMENT:

The data is related to direct marketing campaigns (phone calls) 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 (variable y).

INTRODUCTION:

Given the data of Direct marketing campaigns of Portuguese bank company and these marketing campaigns are primarily focused on phone calls. These phone calls are to check whether the client is subscribed to the short-term-deposit of the particular bank.

DATA SUMMARY:

The Provided data has around 45000 observations client data Portuguese banking corporations.

The dataset contains 45211 rows and 17 columns.

- Age (Numeric)
- Job: type of job (Categorical: Admin, Blue-Collar, Entrepreneur, Housemaid, Management, Retired, self-employed, Services, Student, Technician, Unemployed, Unknown)
- Martial: Marital Status (Categorical: Divorced, Single, Unknown)
- Education: (Categorical: Basic.4y , Basic.6y , Basic.9y , High school , Illiterate , Professional. Course, University. Degree , Unknown)
- Default: Has credit in default? (Categorical: No, yes, Unknown)

- Housing: Has house loan? (Categorical: Yes, No, Unknown)
- Loan: Has personal loan? (Categorical: Yes, No, Unknown)

Related with last contact of current campaign:

- Contact: Communication type
- Month: last contact month of year
- Day of week: last contacted day of the week
- Duration: Last contact duration, in seconds. Important note: this attribute highly affects the output target. Yet, the duration is not known before a call is performed. Also, after the end of call y is obviously known.

Other attributes:

- Campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
- Pdays: number of days that passed by after the client was last contacted from a previous campaign (999 means client was not previously contacted)
- Previous: number of contacts performed before this campaign and for this client
- Poutcome: Outcome of the previous marketing campaign

Output variable (desired target):

- Y – Has the client subscribed a term deposit? (Binary Yes, No)

STEPS INVOLVED:

Exploratory Data Analysis:

We have performed several EDA techniques to clearly check how each and every feature is behaving with respect to the dependent feature y. We have also checked for null values, duplicate values luckily there were none. We have visualised several plots which gives several insights of all the features and we can see which features are more important.

Encoding Categorical Features:

There were totally 17 features out of those 10 were categorical, we have used Label Encoder to convert object data type to numerical data type.

VIF (Variance Inflation Factor):

It is a measure of multicollinearity between features, it describes how much a feature is dependent on other features. Here we have done VIF on the features and all the features have good score there is no much collinearity.

SMOTE (Synthetic Minority Oversampling Technique):

It is an oversampling technique for imbalanced datasets. Unlike other oversampling techniques SMOTE is an advanced method, this creates synthetic data points which are not duplicate data points but are near points to the original data points. Here we have performed this technique on the data and fitted models on the data.

Training and Fitting Models:

We have created several classification models and Boosting techniques, there were different outcomes from each model.

The following are the models created:

Classification Models:

- Logistic Regression
- SVC (Support Vector Classifier)
- Gaussian Naïve Bayes
- Random Forest Classifier
- KNN

Boosting Techniques:

- AdaBoost Classifier
- Gradient Boost Classifier
- XGBoost Classifier

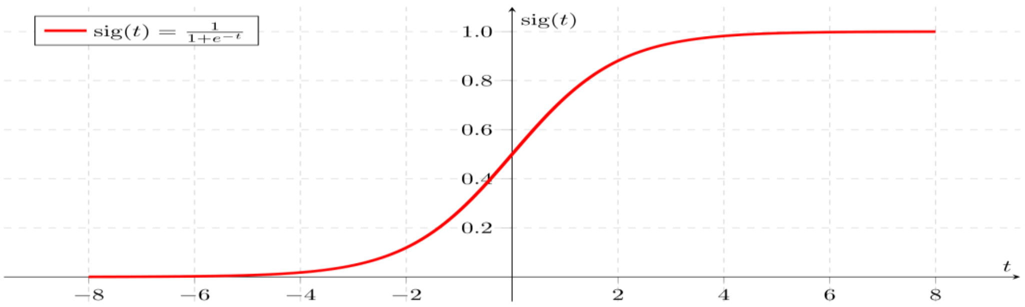
Tuning Hyperparameters for better accuracy:

Tuning hyperparameters are very important for model building especially for tree based algorithms and boosting techniques. This helps us to avoid overfitting and helps to get better accuracy.

Algorithms:

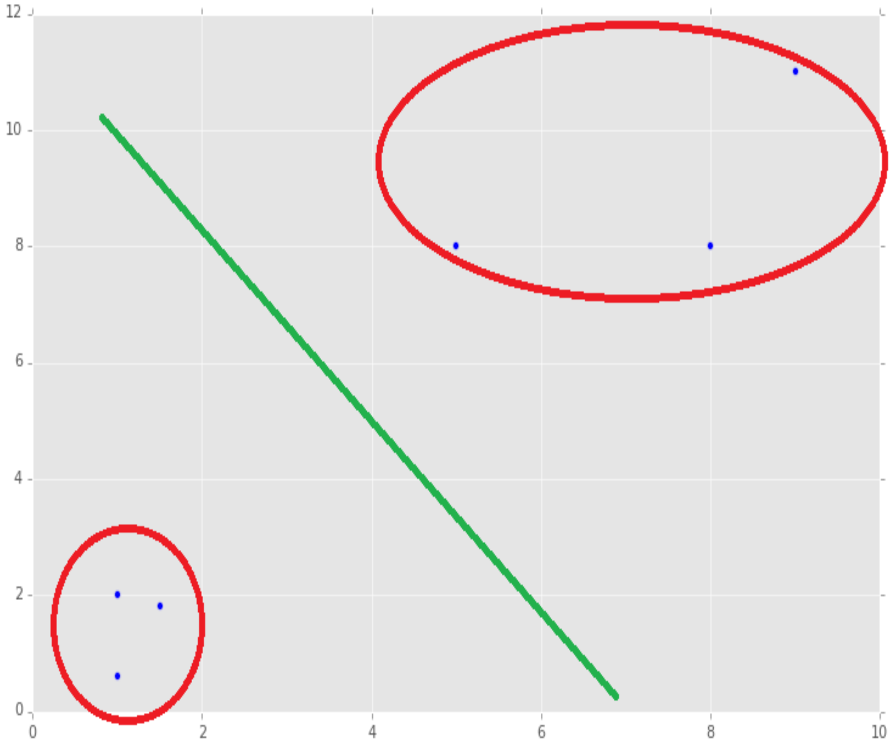
Logistic Regression:

Logistic regression despite its name, it is a classification model. Which is widely used in the industry for classification models. It is a very simple model which distinguishes true/false, yes/no, so on. This is used for binary classification. It is used extensively for linearly separable data. Logistic Regression uses sigmoid function



SVC (Support vector classifier):

Objective of the linear svc model is to fit the data which we provide, which returns a “best fit” hyperplane which divides the data. After getting the hyperplane, we can feed some features to the classifier to see what the predicted class is. This model is suitable for several applications.



Here we can see that two groups of data is divided using hyperplane.

Gaussian Naïve Bayes:

Naïve Bayes makes the assumption of features that they are independent. It means that we are still assuming class-specific covariance matrices, but the covariance matrices are diagonal matrices. It is due to the assumption that the features are independent.

Given a training dataset of N inputs variables x with corresponding target variables t, Naïve bayes assumes that the class conditional densities are normally distributed.

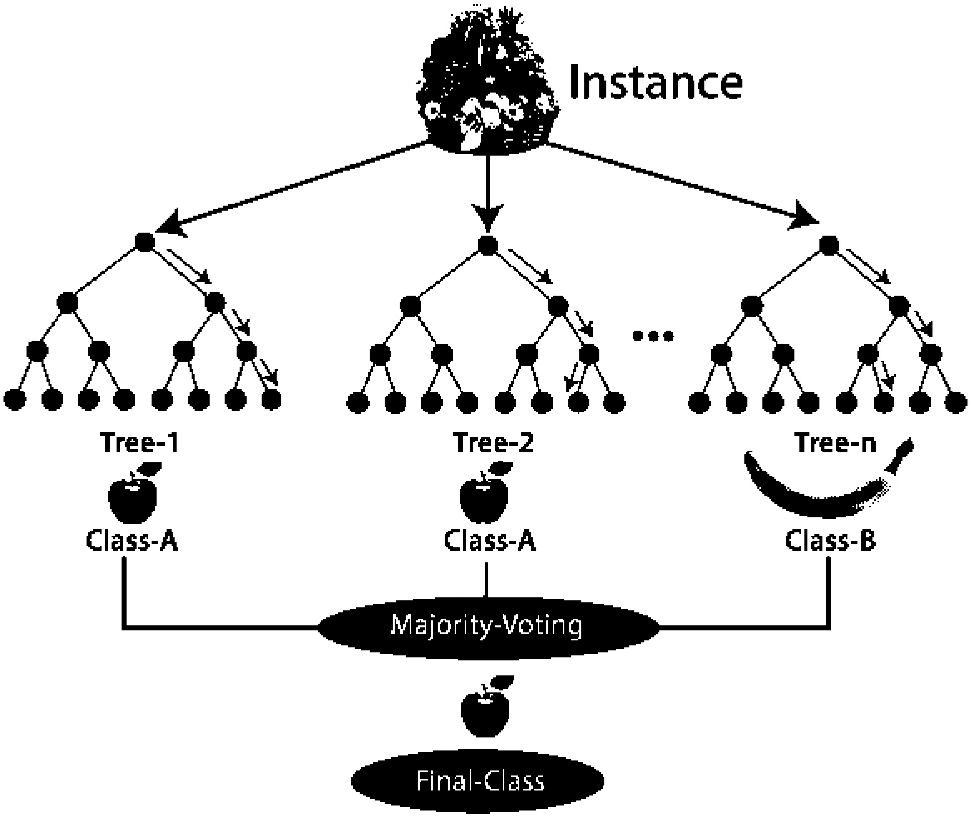
$$P(\mathbf{x} \mid t = c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) = N\left(\mathbf{x} \mid \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right),$$

Where µ is the class specific mean vector and ∑ is the class specific covariance matrix.

Random Forest Classifier:

Random Forest classifier the name suggests it contains a number of decision trees on various subsets of the given dataset and takes the average to improve the accuracy of predictions of the dataset. Instead of relying on one decision tree the random forest uses average of all the decision trees based on majority voting.

The greater the number of trees in the model leads to more accuracy of the model. It also prevents overfitting of the model.

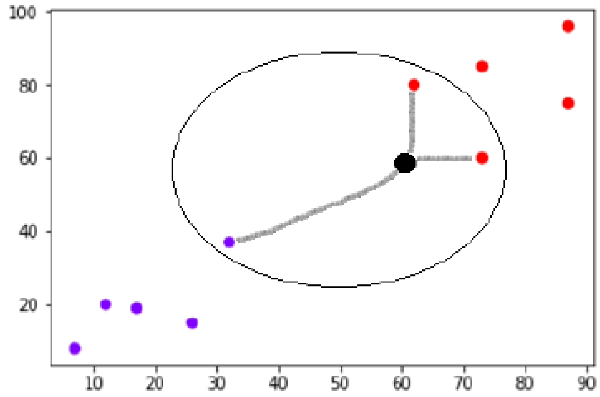


KNN:

It is supervised machine learning algorithm which can be used for both classification and regression problems. Main purpose of knn is for classification problems. The two properties of knn would define well-

- Lazy learning algorithm: Knn is a lazy learning algorithm because it uses all the data for training while classification.
- Non-parametric learning algorithm: It doesn't assume anything about the underlying data.

It uses 'feature similarity' to predict the values of new data points. The new data point will be assigned a value based on how closely it matches the points in the training set.



Above we can see that black dot is new data point and there are two classes as the new point is closer to red points so the new data point will be assigned to red data points.

AdaBoost Classifier:

It is a classifier, which is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases. The core principle of adaboost is to fit a sequence of weak learners.

GradientBoosting Classifier:

It is a group of machine learning algorithm that combine many weak learning models together to create a strong predictive model. It uses decision tree algorithm mostly for doing gradient bosting. It depends on loss function. It has other two necessary parts: a week learner and an additive component. It uses regression trees for the weak learners and this output real values. The additive component of gradient boosting model is added to the model overtime, when this occurs the existing trees aren't manipulated their values remain fixed.

XGBoosting:

It stands for Extreme Gradient Boosting. It is an implementation of Gradient Boosted decision trees. In this algorithm decision trees are created sequentially, where weights play an important role in XgBoost. Weights are assigned to all the independent variables. These are fed to decision tree and the output is the fed to second decision tree. This process repeats for total decision trees assigned. It can work on regression, classification, ranking and user defined prediction problems.

Conclusion:

The Data related to marketing campaign of a Portuguese bank company has been successfully visualized and analysed. After drawing all the possible insights, the data has been worked on pre-processing to undergo all the possible machine learning classification models after carefully fine tuning it.

Metrics of the classification models with normal sampling of dependant variable :

- Logistic Regression – Accuracy : 89%
- SVM Classification – Accuracy : 89.9%
- Gaussian Naive Bayes – Acuuracy : 83.9%
- Random Forest Classifier – Accuracy : 90%
- KNN Classifier -- Accuracy : 89.7%

Boosting Methods :

- ADA Boost Classifier – Accuracy : 90%
- Gradient Boost Classifier – Accuracy : 90.6%
- XG Boosting Classifier – Accuracy : 90.5%

Metrics of the classification models after over-sampling the dependant variable :

- Logistic Regression – Accuracy : 86.2%
- SVM Classification – Accuracy : 89%
- Gaussian Naive Bayes – Acuuracy : 74%
- Random Forest Classifier – Accuracy : 88.4%
- KNN Classifier -- Accuracy : 88.7%
- Boosting Methods :
- ADA Boost Classifier – Accuracy : 62.5%
- Gradient Boost Classifier – Accuracy : 91.5%
- XG Boosting Classifier – Accuracy : 93%

After closely observing all the performed classification models and we can clearly see the over-sampled XG Boosting classifier gave the best results and the predicted values were on par on the test dataset.

References:

1. Medium
2. GeeksforGeeks
3. Analytics Vidhya