Evaluating marketing campaign success prediction using classification

Boris Chan 213574629

Table of Contents

1. Introduction

Marketing campaigns are essential to business success. Marketing allows for a business to engage, sell, grow and learn from customers. This paper will examine and analyze marketing campaign data from a Portuguese banking institution. Classification will be used as the main data mining technique for conducting data analysis. Classification is a predictive model based on supervised learning. The classification algorithms used in the analysis are: logistic regression, naive bayes classifier, decision tree, and random forest classifier.

These algorithms will make data-driven predictions through building a mathematical model from training data and testing data. The training dataset is a sample of the dataset used to fit the model, and the testing dataset is a sample of the dataset used to provide an unbiased evaluation of the model. For evaluating the performance of a single algorithms/models, methods such as cross-validation and bootstrap will be used. For comparing models, the use of ROC curves and confusion matrix can be employed.

1.1 Problem definition/formalization

Given a dataset containing customer information, can we build a model to predict whether a new customer will subscribe to the campaign.

2. Dataset

The dataset is related to direct marketing campaigns of a Portuguese banking institution. The dataset was sourced from the UCI Machine Learning Repository, which retrieved it from the study conducted by Moro et. al [7]. The original dataset contained a large number of features (150), and using knowledge from domain experts they reduced the number of features down to 22 through a series of interviews and question analysis [7].

The dataset contains 41,180 tuples with 20 attributes. Most of the attributes are categorical, and therefore need to be encoded in the data preprocessing step. The attributes describe information about the client, the campaign (that the client currently is in), the economic context, as well as additional information from the previous campaign. The client and campaign related attributes are fairly clear, e.g. age, education, job, contact, however the economic context attributes might need elaboration and are described in the following table:

Table 2.1 - Social and economic context attributes (numerical)

Attribute	Description
emp.var.rate	Employment variation rate (quarterly)
cons.price.idx	Consumer price index (monthly)
cons.conf.idx	Consumer confidence index
euribor3m	Euribor 3 month rate (European banks interest rate)
nr.employed	No. of employees (quarterly)

The dataset also contains the binary attribute ‘y’. As this paper focuses on using classification techniques on the dataset, the ‘y’ attribute will be used as the class attribute when building the models. The dataset will be split into a training set and testing set with an approximately 70/30 split.

3. Data Preprocessing

The categorical attributes need to be encoded and transformed into dummy variables, for example education = { ‘high.school’, ‘university.degree’, ‘professional.course’ … } will be encoded to: education = { 1, 2, 3, … }. Some numeric attributes such as ‘Age’ will be broken down into smaller categories. The ages: ages 0-32 were assigned as 1, ages 32-47 assigned as 2, ages 47-70 assigned as 3 and ages greater than 70 were assigned 4. These ranges were chosen due to the distribution of ages from the dataset, as illustrated in the boxplot below:

Figure 3.1 - Boxplot of ‘age’ attribute

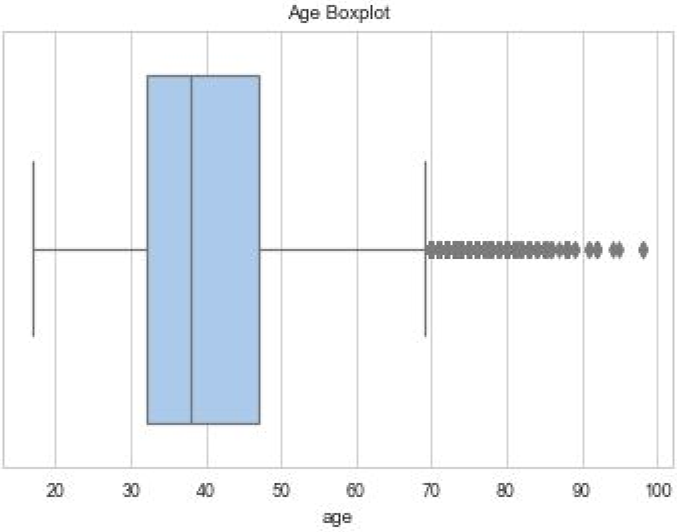


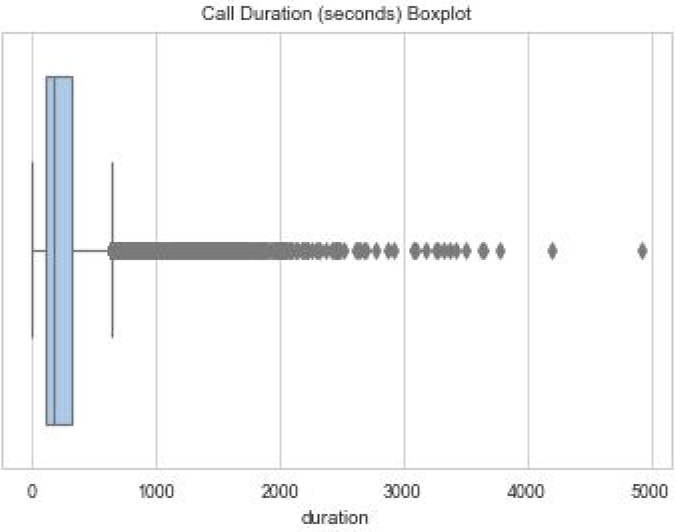
Table 3.1 - Five-number summary of attribute Age.

Stat	Value
Min	17
q_1	32
Median	38
q_3	47
Max	98
$q_3 + \text{IQR } (q_3 - q_1)$	69.5

Table 3.1 illustrates the five-number summary of Age. Data points that that are below Q1 - 1.5*IQR = 9.5 and Q3 + 1.5*IQR = 69.5 can be considered as outliers. Therefore, any age above 69.5 will be assigned a value of 4. Similarly, ages assigned a value of 3 lie between 47 (Q3) and 69.5, ages assigned a value of 2 are between 32 (Q2) and 47 (Q3), and ages assigned a value of 1 are any values less than 32 (Q2).

A similar approach will be assigned to the 'duration' attribute, which describes the duration of phone calls in seconds. Figure 3.2 displays a boxplot of the duration attribute.

Figure 3.2 - Boxplot of Duration



4. Approach

The models were implemented using Python with libraries to assist in data analysis such as Numpy, Pandas, Scikit-learn, etc. Scikit-learn provided the actual implementations for all the models listed in Section 5, as well as a method to split our dataset into training sets and a testing sets.

5. Algorithms

5.1 Logistic Regression

Logistic regression is a method to predict the probability of an outcome that can only have two values (e.g. yes or no). The prediction is based on the use of one or several predictors, which can be numerical or categorical. The logistic regression produces a logistic curve, which values are limited between 0 and 1 [3]. The model can be defined with:

$$P = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Where P is the probability that an observation β_0 is the constant and β_1 the slope defines the steepness of the curve.

5.2 Naive Bayes classifier

The Naive Bayes classifier is based on Bayes' theorem with strong independence assumptions between predictors. Bayes' theorem provides a way of calculating posterior probability P($X \mid C$) from P(C), P(X), and P($C \mid X$). To reduce the computation cost in evaluating P($X \mid C_i$), the naive assumption of class-conditional independence is made, where the naive bayes classifier assumes that the components of X on given class C are independent, and then estimates each of the univariate distributions separately [1, pp. 351-354] [2].

5.3 Decision tree (CART)

Decision tree induction is a method which utilizes a tree model of decisions where each branch represents an outcome of the test, and each leaf holds a class label. A trained decision tree will test a given tuple X (where the class is unknown) using its attributes and a path will be traced to a leaf node which holds the class prediction for X. The CART method produces a binary decision tree, which uses the Gini Index as its splitting rule [1, pp. 332-350]. This can be contrasted with the ID3 method which uses information gain as its attribute selection measure. The decision tree will be used as the base model for Random forests.

$$Gini = 1 - \sum_{i=1}^{c} (p_i)^2$$

5.4 Random Forests

Random forests are an ensemble learning method using classification. Each of the classifiers in the ensemble is a decision tree which results in a “forest”. The individual decision trees are generated using a random selection of attributes at each node to determine the split. Each decision tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. During classification, each tree votes and the most popular class is returned.

6. Experimental Results

The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier [1, p. 365], which can be denoted by:

$$accuracy = \frac{TP + TN}{P + N}$$

It is important to note that accuracy isn’t always the best measure, and we will discuss which measures may be more appropriate in the evaluation section. However, accuracy can currently be used as a general measure to see the performance of the models.

Table 6.1 - Initial accuracy scores from algorithms.

Model	Accuracy (5 s.f.)
Logistic Regression	0.90823
Naive Bayes	0.84656
Decision Tree	0.88459
Random Forests	0.90426

7. Evaluation

7.1 Model selection

A confusion matrix can be used to evaluate the quality of a classifier. The matrix shows the true positives, true negatives, false positives, and false negatives. The measures used to assess a classifier predictive ability are: accuracy, recall, and precision. It should be noted that the accuracy measure might not be reliable when the main class of interest is in the minority [1, p. 386] Precision is a measure of exactness, whereas recall is a measure of completeness [1, p. 386]; these measures can be computed by the following formulas:

$$precision = \frac{TP}{TP + FP} \ , recall = \frac{TP}{TP + FN}$$

Table 7.1 - Confusion matrix (Logistic Regression)

	Predicted 0	**Predicted 1**
Actual 0	10711	257
Actual 1	877	512

Table 7.2 - Confusion matrix (Naive Bayes)

	Predicted 0	**Predicted 1**
Actual 0	9672	1296
Actual 1	600	789

Table 7.3 - Confusion matrix (Decision Tree)

	Predicted 0	**Predicted 1**
Actual 0	10211	757
Actual 1	669	720

Table 7.4 - Confusion matrix (Random Forests)

	Predicted 0	**Predicted 1**
Actual 0	10562	406
Actual 1	777	612

Table 7.5 - Precision, recall and F-1 scores for models (5 s.f.)

Model	Precision	Recall	F Score
Logistic Regression	0.66580	0.36861	0.47451
Naive Bayes	0.37841	0.46542	0.41743
Decision Tree	0.48747	0.51836	0.50244
Random Forests	0.60118	0.44060	0.50851

Precision can be used as a measure for comparing models when the cost of false-positive is high. In predicting marketing campaigns, a false positive means that a prospective client that is not going to purchase (actual negative) has been identified as likely to purchase (predicted positive).

Similarly, recall can also be used as a measure when the cost of false-negative is high. If a prospective client that is going to purchase (actual positive) is predicted as not going to purchase (predicted negative). Both situations can result in negative consequences depending on how the organization uses this information. Table 7.5 illustrates that the Logistic Regression model offered the highest precision measure, whereas the Decision Tree offered the highest recall measure.

Given these scenarios, it can be argued that using recall is a better measure for comparing the models, as the cost of false-negative is high. If a prospective customer has been predicted negative, i.e. not going to purchase, then the organization risks losing potential business.

The f-score measure is the harmonic mean of precision and recall, where it gives equal weight to precision and recall. When a model has perfect precision and recall, the f score would be 1, and when a model has the lowest precision and lowest recall, the f score would be 0 [6]. The results above show that the Random Forests classifier has the highest f-score. The f-score should be considered when scoring a model when an organization values both the precision and recall and their respective implications.

7.2 K-fold cross validation

It is important to note that the accuracy scores are not always the same. This is because the training and testing set are selected at runtime, therefore different samples are selected every time the code is executed and hence affect the model. This is why k-fold cross validation is required.

In k-fold cross validation, the dataset is randomly partitioned into k subsets, $D_1, D_2, \ldots, D_k$. Training and testing is done *k* times. In iteration *i*, subset D_i is reserved as the test set, and the remaining subsets are used to train the model. For instance, when *i* = 1, $D_2, \ldots, D_k$ will be used to train the model and D_1 will be used as the testing set; then when *i* = 2, $D_1, D_3, \ldots, D_k$ will be used to train the model and D_2 will be used as the testing set, etc. [1, p. 370].

This paper will use stratified cross-validation, where the folds are stratified to ensure that each fold is a good representative of the whole. For example, in a binary classification problem where each class consists of 50% of the data, it is best to arrange the data so that every fold follows this distribution [4]. A study conducted by Kohavi concludes that using 10-fold stratified cross validation is the best method to use for model selection, in terms of of bias and variance when compared to regular cross-validation [5].

Table 7.6 - Sample mean accuracy for 10-fold cross validation (5 s.f.)

Model	**Accuracy**	**Error Rate**	**Margin of Error**
Logistic Regression	0.85796	0.14203	0.080084
Naive Bayes	0.78564	0.21435	0.16406
Decision Tree	0.58977	0.41023	0.17957
Random Forests	0.64302	0.35698	0.20319

The results pose the following question: why did the random forests classifier perform so poorly? An ensemble method should be expected to perform similar if not better than a base learner. The results of the cross validation on the random forests model are illustrated in Figure 7.1:

Figure 7.1 - Accuracy scores of Random Forests from 10-fold cross validation.

Mean accuracy of Random Forests				
0.88759408	0.86404467	0.75795096	0.85214858	0.83806749
0.78368536	0.37387715	0.14396698	0.70665372	0.22219524

The accuracy went as low as 0.14 for one of the cases in the cross-validation. Stratified k-fold cross validation was used, so the case of having an unbalanced distribution of training and testing samples would be highly unlikely. A study conducted by Kirasch conclude that when increasing the variance in explanatory and noise variables, logistic regression consistently performed with a higher overall accuracy as compared to random forest [8]. The decision tree model performed worse as well, which suggests that a possible solution is to perform pruning. Tree pruning algorithms attempt to improve accuracy by removing tree branches reflecting noise in the data [1, p. 385]. An investigation of why the random forests and decision tree performed so poorly require a more thorough analysis.

Nevertheless, the mean error rates calculated in the cross-validations may be considered as different independent samples from a probability distribution, where they follow a t-distribution with $k-1$ degrees of freedom (where $k=10$ from the 10-fold cross validation). We can conduct a test of statistical significance to know if the difference between 2 error rates is attributed to chance.

Our hypothesis is that two models are the same (the difference in mean standard error is 0). If we reject this hypothesis (null hypothesis), then we can conclude that the difference between the two models is statistically significant, in which case we can select the model with the lower error rate [1, p. 372]. Let us compare Logistic Regression and Naive Bayes classifiers (as they have the lowest error rates).

H_0 : Logistic Regression has the same performance as Naive Bayes.
H_a : The models have a difference in performance.

$$t = \frac{err(M_1) - err(M_2)}{\sqrt{var(M_1 - M_2) / k}}$$

The t-test for the mean accuracies from the 10-fold cross validation of the Logistic Regression and Naive Bayes resulted in a t-statistic of: 0.89607 and a p-value of 0.38204. From the t-table shows that the critical value 2.262 (at $k - 1$ degrees of freedom). Our t-statistic is less than the critical value, and therefore we fail to reject the null hypothesis. From this test we can say that there is no significant difference between the two models.

The reasons why the t-test failed could be due to the fact that the variance (or noise) from the dataset is too high. High variance was also suggested by the poor performance of the decision tree and random forests model. If this study was to be continued further, methods to reduce noise and outliers such as DBSCAN will need to be done during the data preprocessing phase.

7.3 ROC curves

Receiver operating characteristic (ROC) curves are a useful visual tool for comparing classification models. The curve for a given model shows the trade-off between true positive rate and false positive rate, where the x-axis represents the false positive rate and the y-axis represents the true positive rate [1, p. 374]. The formula for both rates can be denoted by:

$$TPR = \frac{TP}{P} \quad , \quad FPR = \frac{FP}{N}$$

The ROC curves for the models used in this paper are displayed in Figure 7.2 - 7.7

Figure 7.2 - Logistic Regression

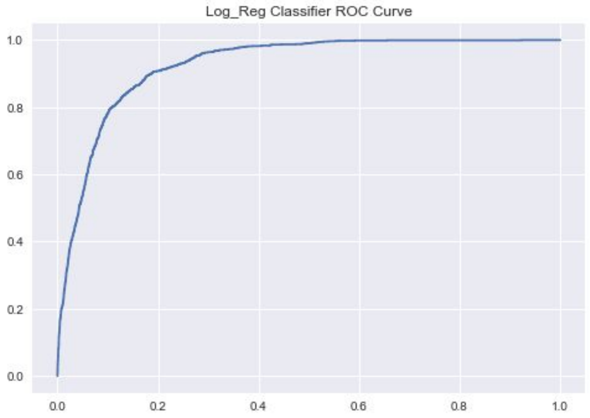


Figure 7.3 - Naive Bayes

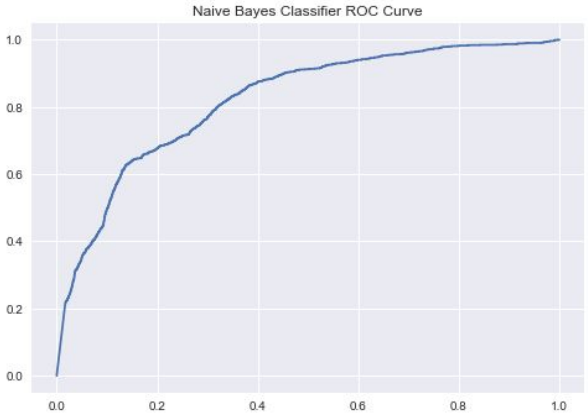


Figure 7.4 - Decision Tree

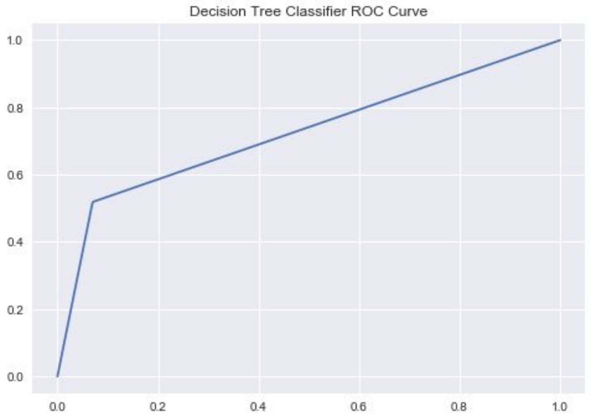
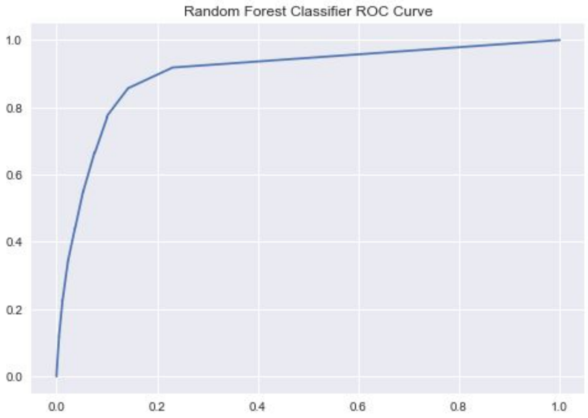


Figure 7.5 - Random Forests



The diagonal line (y = x) indicates that for every true positive, there is a false positive. The closer the ROC curve is to the diagonal line, the less accurate the model is. To accuracy of a given model can be assessed by measuring the area under the curve (AUC). The closer the area is to 0.5, the less accurate the corresponding model is. A model with perfect accuracy will have an area of 1.0 [1, p. 377). The Logistic Regression model had the highest AUC score followed by the Random Forests classifier. The AUC scores for each model are illustrated in the following table:

Table 7.8 - AUC values

Model	**AUC**
Logistic Regression	0.92556
Naive Bayes	0.82047
Decision Tree	0.72436
Random Forests	0.90012

8. Conclusion

In summary, our findings showed that the best model depends on the measure that is most valued by the organization. Precision, recall, and f-scores are popular measures that can be used depending on the situation, however it was argued that for predicting marketing campaign success, the cost of false-negative is high, therefore a high recall score should be reflected in the chosen model. Table 7.5 illustrates that the Decision Tree model had the highest recall, however from further evaluation we concluded that the Decision Tree and Random Forest models should not be used without additional data preprocessing and cleaning steps.

In Section 7.2, the results of the k-fold cross-validation showed that the Decision Tree and Random Forests performed poorly. Furthermore, we failed to reject the null hypothesis, that is to say: there was no statistically significant difference between Logistic Regression and Naive Bayes model. All these indications point to the fact that the dataset has too high variance and noise, and therefore need to be preprocessed further using methods such as DBSCAN. Overall, given the current dataset, the Logistic Regression model had the lowest error rate and highest ROC score. Regardless, the reliability of these scores are questionable due to the state of the dataset and therefore cannot be fully recommended.

9. References

[1] J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann; 3 edition (July 6, 2011), 2011.

[2] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, and etal., “Top 10 algorithms in data mining,” Knowledge and Information Systems, vol. 14, no. 1, pp. 1–37, 2007.

[3] "Logistic Regression", Saedsayad.com, 2019. [Online]. Available: https://www.saedsayad.com/logistic_regression.htm. [Accessed: 27- Mar- 2019].

[4] "Understanding stratified cross-validation", Cross Validated, 2019. [Online]. Available: https://stats.stackexchange.com/questions/49540/understanding-stratified-cross-validation. [Accessed: 27- Mar- 2019].

[5] R. Kohavi, A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. 1995.

[6] “F-Score Definition,” DeepAI. [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/f-score. [Accessed: 27-Mar-2019].

[7] Moro, Sérgio & Cortez, Paulo & Rita, Paulo. (2014). A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems. 62. 10.1016/j.dss.2014.03.001.

[8] Kirasich, Kaitlin; Smith, Trace; and Sadler, Bivin (2018) "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets," SMU Data Science Review: Vol. 1 : No. 3 , Article 9.