School of Continuing Studies

Big Data Management Systems and Tools

Predicting The Success of Bank Telemarketing Strategies Using MLlib

Term Project

December 3, 2019

Authors:

Sakshi Sharma

TABLE OF CONTENTS

1.0 Introduction

Technology has enabled focused efforts in the marketing campaigns across various businesses by focusing on maximizing customer lifetime value by evaluating existing information and customer engagement metrics. This has led to the development of Decision Support Systems (DSS) that drive faster, smart decisions based on data rather than subjective, biased and personal instincts [2]. Implementing DSS provides the right information to the agents who can then focus their efforts to work smarter and meeting operational KPIs.

These systems are relevant for many verticals including healthcare, finance, weather prediction, call and chat centers, desktop apps, information kiosks and more.

Telemarketing is defined in Section 52.1 of the Competition Act of Canada [1]. It is the practice of using "interactive telephone communications" for the purpose of promoting directly or indirectly any product or business interest. Telemarketing campaigns are typically led through call centers. These workplaces are highly stressed environments where agents' performance is closely monitored and their actions have business implications, meeting sales goals is usually one of the many implications[2].

Agent's access to DSS is critical in meeting sales campaign goals and is integral to enhance business and aim at meeting specific targets.

In real life, the telemarketing calls may be outbound or inbound, depending on which side triggered the contact, client or call center agent. The goal of the agent is to primarily provide relevant information to the client and serve the client's best interests. The end result for the business is the conversion ratio which is the number of sales made during any given campaign.

Traditionally, the DSS models have been built by humans and then developed into a workflow. But the relatively new integration fo AI into DSS has created more sophisticated, problem-oriented and intelligent decision-support systems (IDSS) that can understand a wide range of inputs and select the next best course of action.

The intention of the term project is to build IDSS that can be applied specifically to the area of targeted marketing that will allow the agents to focus on a particular group of society/existing clients to maximize the chances of making a sale. The IDSS can predict the result of a phone call to sell long term deposits.

1.1 Scope

The data used for this study is the bank marketing dataset posted on UCI Machine Learning Repository. This is something of personal interest to me as I have previously worked in the banking industry and I was motivated to explore the implementation of big data and machine learning principles on decision support systems.

The dataset includes a large dataset from a Portuguese banking institution. The data was collected during a direct marketing campaign from 2008 to 2013. The data includes more commonly used bank client and product attributes. The features available in the data have already been analyzed and shortlisted from 150 features [3].

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 [3].

1.2 Objectives

In particular, classification is the most common DM task and the goal is to build a data-driven model that learns an unknown underlying function that maps several input variables, which characterize an item (e.g., bank client), with one labeled output target (e.g., type of bank deposit sell: “failure” or “success”). The classification goal is to predict if the client will subscribe (yes/no) to a term deposit [4].

The model development will be done using Scala and MLlib library and associated functions.

2.0 Data Preparation

The data was available as a comma-separated file with ‘;’ used as a delimiter. The file was imported into data bricks for further evaluation.

2.1 Data Quality

This study considers real data collected from a Portuguese retail bank, from May 2008 to June 2013, in a total of 52,944 phone contacts [3]. Each record included the output target, the contact outcome ({“failure”, “success”}), and candidate input features. The features are summarized in Table 1.

Table 1 - Summary of Attributes

Attribute	Attribute Type	Details
BANK CLIENT DATA		
age	numeric	
job	type of job, categorical	"Admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services"
marital	marital status, categorical	"married","divorced","single"; note: "divorced" means divorced or widowed
education	categorical	"unknown","secondary","primary","tertiary"
housing	has a housing loan? categorical	'no','yes','unknown’
loan	has a personal loan? categorical	'no','yes','unknown’
RELATED WITH THE LAST CONTACT OF THE CURRENT CAMPAIGN		
contact	contact communication type, categorical	Cellular','telephone', 'unknown’
month	last contact month of year, categorical	Jan to Dec
date_of_month	last contact day of the month	numeric
OTHER ATTRIBUTES		

campaign	number of contacts performed during this campaign and for this client, numeric	
pdays	number of days that passed by after the client was last contacted from a previous campaign, numeric	
previous	number of contacts performed before this campaign and for this client, numeric	
poutcome	outcome of the previous marketing campaign, categorical	'failure','nonexistent','success'
y	Outcome, categorical	‘Yes’, ‘no’

The “Duration” field is not used as one of the features in the development of IDSS as this information is usually not available before a call is made. In addition, after the end of the call, y is obviously known. This is discarded in the hopes of developing a realistic model.

2.2 Challenges

The data did not have any missing attribute values. However, the outcome (yes or no) was saved as a string. The classification algorithms required the conversion of text to string. As such, “yes” was changed to “1” and “no” was saved as “0” for further analysis.In addition, the data is imbalanced as shown in Figure 2-1.

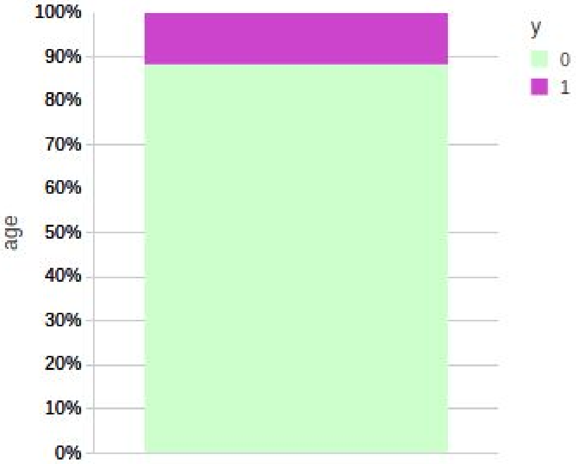

Figure 2-1 - Illustration of Unbalanced Dataset for the Outcome Field

2.3 Feature Creation

No additional features were created for the analysis.

3.0 Data Exploration

Figure 3-1 shows that there are distinct demographics that have successfully subscribed to a term deposit. This suggests that the recorded demographic affects the final outcome. In addition, there is a clear demarcation of groups who are more willing and susceptible to marketing campaigns. Therefore the data can be used to develop intelligent DSS.

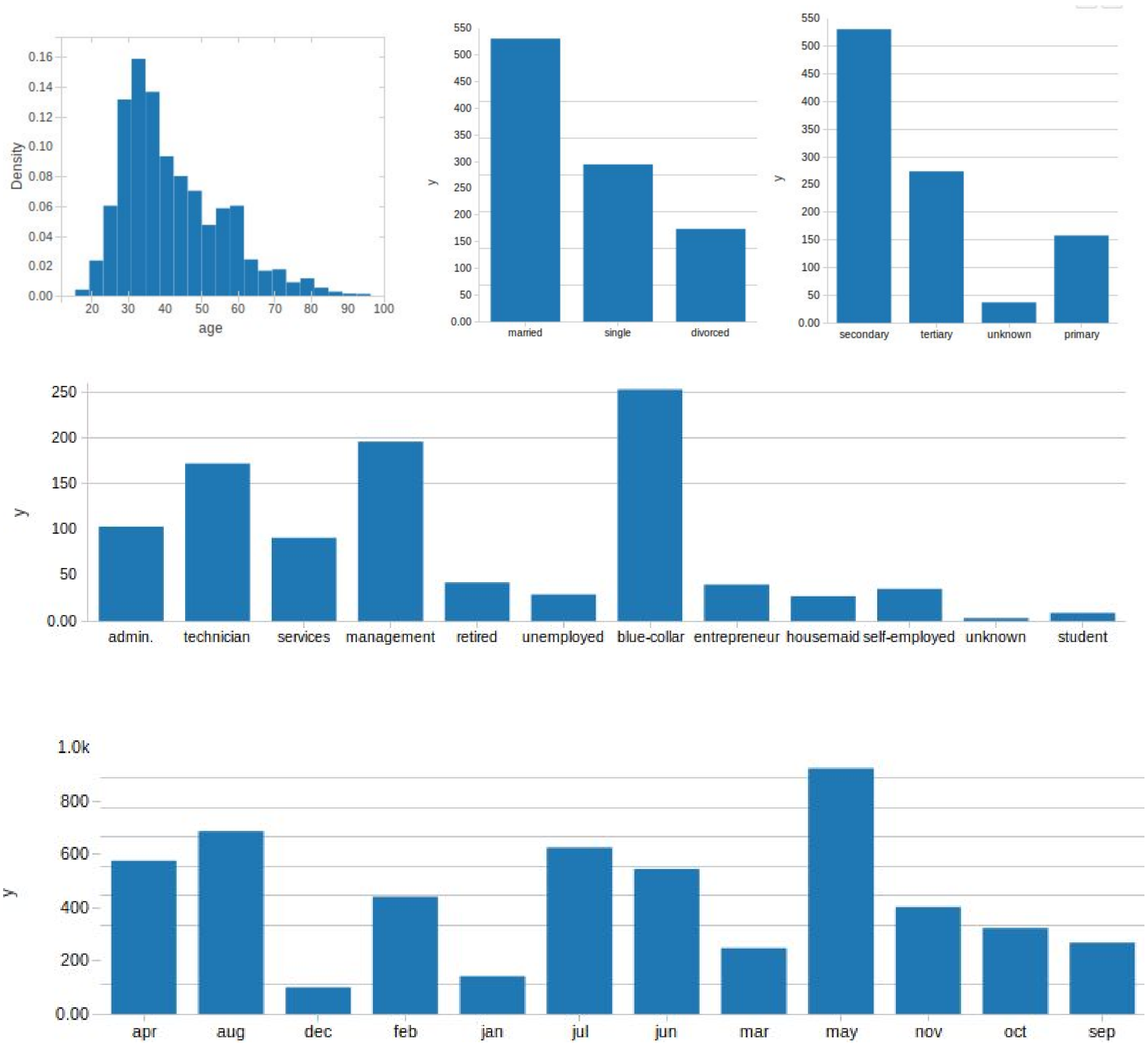

Figure 3-1 - Sample Demographics of Successful Telemarketing Calls (Age, marital, education, job, month)

4.0 Model Development and Evaluation

4.1 Methodology

The following approach was used to prepare the data for model development, evaluation and selection.

1. Apply ‘StringIndexer’ to the categorical features.
2. Apply ‘OneHotEncoderEstimator’ to the indexed categorical features.
3. Select the meaningful features were assembled using ‘VectorAssembler’
4. Apply ‘StandardScaler’ to transform the dataset of vector rows, normalizing each feature to have a unit standard deviation.
5. Finally, the data set was split into test and train using ‘random split’

A pipeline was created for each model that included model selection (a.k.a hyperparameter tuning) using cross-validation on the training dataset. BinaryClassificationEvaluator and its metric 'areaUnderROC' was used during cross-validation.

The tuned and selected best model was used to predict the outcome of the test data. BinaryClassificationEvaluator and its metric 'areaUnderROC' was used during model evaluation.

The following models are considered for the development of IDSS [5] -

1. Binomial Logistic Regression
    a. Logistic regression is widely used to predict a binary response.
2. Random Forest Classifier (ensembles of decision trees)
    a. Random forests combine many decision trees in order to reduce the risk of overfitting
3. Linear Support Vector Machine
    a. A support vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification

Hyperparameter tuning was performed for each model using cross-validation (3 folds) for all the three models. The best model for all models was selected after this process and the test data was used to predict the outcome based on the three different modeling approaches.

The data transformation pipeline is shown in Figure 4-1 and the corresponding model tuning and evaluation pipeline are shown in Figure 4-2.

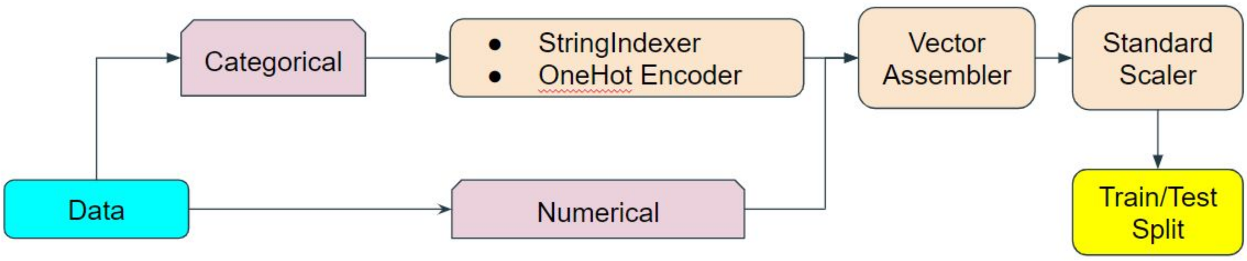

Figure 4-1 - Data Transformation Pipeline

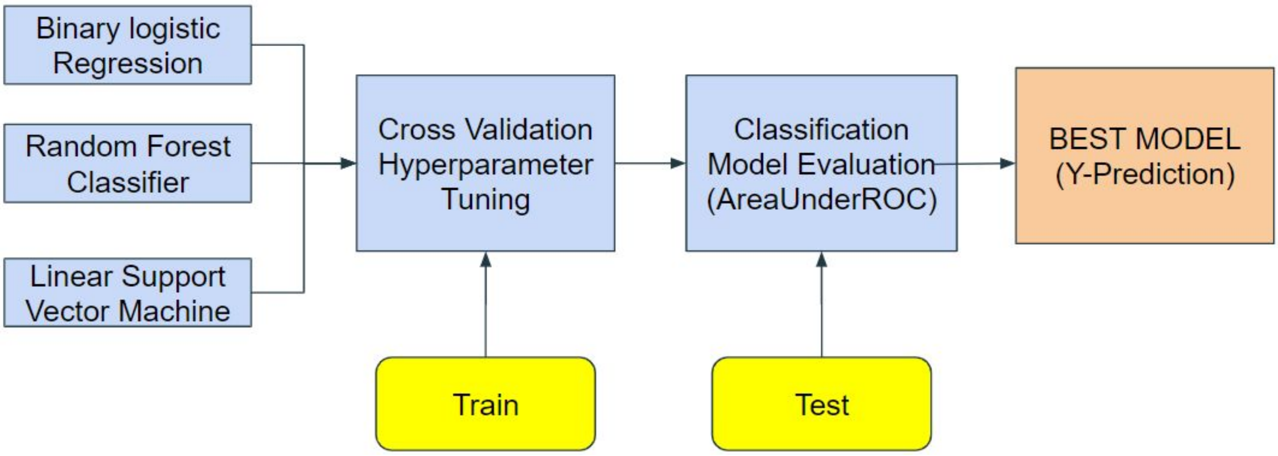
Binary logistic Regression
Random Forest Classifier
Linear Support Vector Machine
Cross Validation Hyperparameter Tuning
Classification Model Evaluation (AreaUnderROC)
BEST MODEL (Y-Prediction)
Train
Test

Figure 4-2 - Model Tuning and Evaluation Pipeline

4.2 Discussion

For the project, as mentioned above, the accuracy of the model is measured by the area under the ROC (Receiver Operating Characteristics) curve. It tells how much the model is capable of distinguishing between classes. Higher the AUROC, the better the model is at predicting 0s as 0s and 1s as 1s. An area of 1 represents a perfect test; an area of .5 represents a worthless test. A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system [6] - 0.90-1 = excellent (A), 0.80-0.90 = good (B), 0.70-0.80 = fair (C), 0.60-0.70 = poor (D) and 0.50-0.60 = fail (F).

The results from the cross-validation tests on the optimized hyperparameters (marked as the best model) are shown in Table 2. The best model was then used to predict the outcome from the features. The performance of the model is also summarized in Table 2.

Table 2 - Summary of Model Performance

Model	CrossValidation Score (Best Model)	Model Score on Test Data (Best Model)	Model Rating (A..F)
Binomial Logistic Regression	0.7572	0.7637	C, Fair
Random Forest Classifier	0.7833	0.7877	C, Fair
Linear Support Vector Machine	0.6717	0.6649	D, Poor

The ratio of correct and wrong predictions was also calculated in order to assist in model evaluation. These are calculated as follows:

- Correct Prediction Ratio = right prediction / total instances
- Wrong Prediction Ratio = wrong prediction / total instances

Table 3 - Ratio of Correct and Wrong Prediction

Model	Correct Prediction Ratio	Wrong Prediction Ratio
Binomial Logistic Regression	0.8885	0.1115
Random Forest Classifier	0.8922	0.1077
Linear Support Vector Machine	0.8841	0.1158

As seen from the various model evaluation metrics, Random Forest Classifier obtains the best classification.

5.0 Conclusions

From the model performance, binomial logistic regression and random forest classifier models have the best probability of correctly predicting the target demographic group. These models can be used by the bank for developing intelligent Decision Support Systems.

This study clearly demonstrates that the machine learning approach can be successfully implemented in a Big Data platform, specifically Scala to achieve the classification goal and predict if the client will subscribe (yes/no) to a term deposit based on a given dataset.

A possible marketing strategy from the outcomes may consist of the following steps to optimize resource utilization and maximize the probability of achieving campaign success.

1. Focus efforts on approaching potential targets as identified by Random Forest Classifier Model
2. Focus efforts on approaching potential targets as identified by Linear SVM Model
3. If resources are still available and the marketing goal has not been met, approach the rest of the demographics

6.0 Future Work

1. Use stratified sample split so that the proportion of successful and unsuccessful call remains the same in the test and train data test for model development.
2. Explore ensemble models to improve model accuracy.

7.0 Appendix

7.1 References

1. Section 52.1 of the Competition Act, Accessed: November 22, 2019, https://www.competitionbureau.gc.ca/eic/site/cb-bc.nsf/eng/03123.html#intro
2. The Future of Intelligent Decision Support Systems in Contact Centers, Accessed: November 22, 2019, https://techsee.me/blog/decision-support-system
3. 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
4. Bank Marketing Data Set, UCI Machine Learning Repository, Accessed: November 22, 2019, http://archive.ics.uci.edu/ml/datasets/Bank+Marketing
5. Classification and regression, Accessed: November 22, 2019, https://spark.apache.org/docs/latest/ml-classification-regression.html
6. The Area Under an ROC Curve, Accessed: Nov 24, 2019http://gim.unmc.edu/dxtests/roc3.htm