Machine Learning Engineer Nanodegree

Capstone Proposal

Himanshu Sewalkar
April , 2018

Proposal

Domain Background

Direct marketing is a form of advertising where organizations communicate directly to customers through a variety of media including cell phone text messaging, email, websites, online adverts, database marketing, fliers, catalog distribution, promotional letters and targeted television, newspaper and magazine advertisements as well as outdoor advertising. The common form of direct marketing is telemarketing, in which marketers contact customers by phone. The primary benefit to businesses is increased lead generation, which helps businesses increase sales volume and customer base. The most successful telemarketing service providers focus on generating more "qualified" leads that have a higher probability of getting converted into actual sales.

If we consider banking sector, then we can notice that post 2008 economic crisis there is greater pressure on banks to increase profits while reducing costs. Thus, optimization of direct marketing effort under telemarketing has become a key concern. Direct marketing within banking sector is a method where customers are directly informed about banking products, such as credit cards, new savings account types etc. which are analyzed and selected per customer's characteristics and contacted over a communication channel like phone calls.

In direct marketing, the return rate and the effectiveness of the campaigns can be measured using the responses of customers, and improvements on the campaigns can also be made.

Problem Statement

A Portuguese bank had experienced a revenue decline. It was found the root cause is that bank's customers are not depositing as frequently as before.

A term deposits benefits a bank in the following way:

- It can invest the deposits in those financial products which lead to higher gain; so as to make a profit which is higher than the cost associated with deposits.
- Secondly, more the term deposit customers higher is the probability that a bank can upsell or cross sell other products to further increase revenues.

Hence, the Portuguese bank would like to identify existing clients that have higher chance to subscribe for a term deposit and thus focus marketing effort on such clients.

Datasets and Inputs

The dataset we strive to use in the capstone project is from the UC Irvine Machine Learning Repository and holds information related to a direct marketing campaign of the previously described Portuguese bank.

It was obtained by downloading `bank-additional-full.csv` (contained in `bank-additional.zip`) from https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

Shape of the Dataset

The table contains 41,188 rows (i.e., tuples describing telemarketing interaction) and 21 columns (i.e., features of each interaction).

Input Variables

Customer Data

#	Column Name	Description	Type	Values
1	Age	Customer's age	Numerical	
2	Job	Customer's Profession	Categorical	• `admin` • `blue-collar` • `entrepreneur` • `housemaid` • `management` • `retired` • `self-employed` • `services` • `student` • `technician` • `unemployed` • `unknown`
3	Marital	Marital Status	Categorical	• `divorced` • `married` • `single` • `unknown` Note: `divorced` means divorced or widowed
4	Education	Customer's Education Level	Categorical	• `basic.4y` • `basic.6y` • `basic.9y` • `high.school` • `illiterate` • `professional.course` • `university.degree` • `unknown`

5	Default	Indicator if the customer has credit in default	Categorical	• no • yes • unknown
6	Housing	Indicator if the customer has a housing loan	Categorical	• no • yes • unknown
7	Loan	Indicator if the customer has a personal loan	Categorical	• no • yes • unknown

Data related with the last contact of the current campaign

#	Column Name	Description	Type	Values
8	Contact	Contact communication type	Categorical	• cellular • telephone
9	Month	Month that last contact was made	Categorical	• jan • feb • ⋮ • dec
10	Day_of_week	Day that last contact was made	Categorical	• monday • tuesday • Wednesday • Thursday • Friday
11	Duration	Duration of last contact in seconds	Numerical	Note: This attribute highly affects the output target (e.g., if duration=0 then y=no). Yet, the duration is not known before a call is performed. Also, after the end of the call, y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

Other Attributes

#	Column Name	Description	Type	Values
12	Campaign	Number of contacts performed during this campaign for this client (including last contact)	Numerical	
13	Pdays	Number of days since the client was last contacted in a previous campaign	Numerical	**Note:** `999` means client was not previously contacted
14	Previous	Number of contacts performed before this campaign for this client	Numerical	
15	poutcome	Outcome of the previous marketing campaign	Categorical	• `failure` • `nonexistent` • `success`
16	emp.var.rate	Employment variation rate (quarterly indicator)	Numerical	
17	cons.price.idx	Consumer price index - monthly indicator	Numerical	
18	cons.conf.idx	Consumer confidence index - monthly indicator	Numerical	
19	euribor3m	Euribor 3-month rate (daily indicator)	Numerical	
20	nremployed	Number of employees (quarterly indicator)	Numerical	

Output Variable (desired target)

#	Column Name	Description	Type	Values
1	Y	Indicator if the client has subscribed for a term deposit	Binary	• `yes` • `no`

Solution Statement

The output variable (y) is having binary response which means either the customer will subscribe for a term deposit or not. Therefore, I will use a classification approach to predict which customers are more likely to subscribe for term deposits. The methods I will attempt to try are (i) logistic regression, (ii) naïve bayes.

My solution will aim for the following two outcomes:

(i) Determining which variables are important
(ii) Implementing the classifier

to predict if a customer will subscribe for a term deposit.

Benchmark Model

Logistic regression is mainly used in cases where the output is Boolean. So is the case under consideration, the desired target is either a customer subscribes for a term deposit or not. Here, we consider features and outcome Y which can take two values {0,1}. Logistic regression measures the relationship between an output variable Y (categorical) and one or more independent variables, which are usually (but not necessarily) continuous, by using probability scores as the predicted values of the dependent variable. So, logistics regression will be my benchmark model.

Evaluation Metrics

Naive Bayes classifiers have worked very well in many practical situations, famously known spam filtering. Naive Bayes also assumes that the features are conditionally independent. Real data sets are never flawlessly independent but they can be close. In short Naive Bayes has a higher bias but lower variance compared to logistic regression. Therefore, it will be interesting to use Naïve Bayes as solution model and compare that against the benchmark model based on logistics regression.

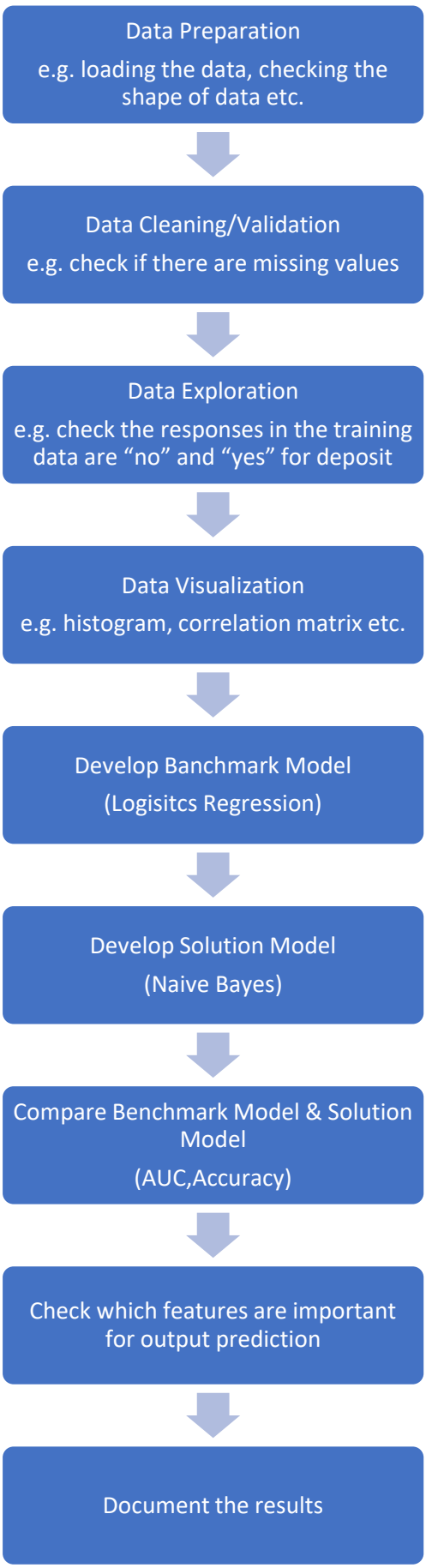
Both Naive Bayes and Logistic regression are linear classifiers, Logistic Regression makes a prediction for the probability using a direct functional form whereas Naive Bayes figures out how the data was generated given the results.

I will use Area Under Curve (AUC) and the accuracy as the two evaluation metrics for the comparison of the models.

	Naïve Bayes	Logistics Regression
AUC	X1	X2
Accuracy (test)	Y1	Y2

Project Design

The workflow of my capstone project analysis and completion will be as depicted in the below flow diagram.


Data Preparation
e.g. loading the data, checking the shape of data etc.
Data Cleaning/Validation
e.g. check if there are missing values
Data Exploration
e.g. check the responses in the training data are “no” and “yes” for deposit
Data Visualization
e.g. histogram, correlation matrix etc.
Develop Banchmark Model
(Logisitcs Regression)
Develop Solution Model
(Naive Bayes)
Compare Benchmark Model & Solution Model
(AUC,Accuracy)
Check which features are important for output prediction
Document the results

References:

https://medium.com/@sangha_deb/naive-bayes-vs-logistic-regression-a319b07a5d4c

http://scikit-learn.org/stable/modules/naive_bayes.html

https://en.wikipedia.org/wiki/Direct_marketing

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

https://pdfs.semanticscholar.org/a911/cbe221347b400d1376330591973bb561ff3a.pdf

https://pdfs.semanticscholar.org/cab8/6052882d126d43f72108c6cb41b295cc8a9e.pdf