Machine Learning Engineer Nanodegree

Capstone Proposal

Armando Perez
December 3, 2018

Proposal

Domain Background

I have about 10 years of professional experience in sales and marketing, and I thought that a project with a domain background in these areas would benefit from my prior experience with them.

Sales and marketing are absolutely vital to most businesses; but businesses have limited resources they can allocate to their marketing efforts. For this reason, they need to know who are the people and or businesses that are most likely to buy. Those who have the best tools and data for predicting potential buyers, will make large amounts of money.

Examples of such profitable businesses are Facebook and Google. Most of their revenue comes from advertising.

Relevant Paper: 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

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

Problem Statement

The classification goal is to predict if the client will subscribe (yes/no) to a term deposit (variable y). The data 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.

Dataset and Inputs

Data Source:

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]

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

Attribute Information:

Input variables:

bank client data:

1. age (numeric)
2. job: type of job (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 (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
5. default: has credit in default? (categorical: 'no', 'yes', 'unknown')
6. housing: has housing loan? (categorical: 'no', 'yes', 'unknown')
7. loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

related with the last contact of the current campaign:

8. contact: contact communication type (categorical: 'cellular', 'telephone')
9. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10. day_of_week: last contact day of the week (categorical: 'mon', 'tue', 'wed', 'thu', 'fri')
11. duration: last contact duration, in seconds (numeric). Important 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:

12. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14. previous: number of contacts performed before this campaign and for this client (numeric)
15. poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')

social and economic context attributes

16. emp.var.rate: employment variation rate - quarterly indicator (numeric)
17. cons.price.idx: consumer price index - monthly indicator (numeric)
18. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
19. euribor3m: euribor 3 month rate - daily indicator (numeric)
20. nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):

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

The classes in my dataset are not balanced (target label 11.2654% yes, 88.7346% no).

I will not do anything to maintain class balances across each subset of the data because I obtained the same ratio as in the main dataset for training and testing set (11.1556% to 88.8444%).

Solution Statement

In this project, I will employ several supervised algorithms to accurately predict if the client will subscribe (yes/no) to a term deposit (variable y). The data 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. I will then choose the best candidate algorithm from preliminary results by their f-score and further optimize this algorithm to best model the data. My goal with this implementation is to construct a model that accurately predicts whether a client will subscribe to a term deposit. This sort of task can arise in a financial sector setting, where organizations are trying to determine who are their more likely future customers. Understanding who are these people can help optimize a financial institution’s marketing efforts by focusing on their more likely customers first.

Benchmark Model

In the paper “A Data-Driven Approach to Predict the Success of Bank Telemarketing”, the researchers achieved 81% classification accuracy on a dataset very similar to our “bank-additional-full.csv” dataset; which through a Lyft analysis their model could select 79% of the positive responders with just 40% of the customers (Moro, 2014, p. 2).

Relevant Paper:

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

Evaluation Metrics

Accuracy as a metric for evaluating a particular model's performance is sometimes appropriate, since we want to accurately determine who our subscribers are. However, accuracy only performs well if the dataset classes are balanced (Afonja, 2017). The classes in our dataset are not balanced (target label 11.2654% yes, 88.7346% no), so accuracy isn’t a viable evaluation metric.

Identifying the clients that did not subscribe to a term deposit as someone who did would be detrimental to us, since we are trying not to waste our marketing dollars on individuals that aren’t likely to subscribe to a term deposit. Therefore, a model's ability to precisely predict those that subscribe is more important than the model's ability to recall those individuals. In particular, when $\beta = 0.5$, more emphasis is placed on precision. This is called the F-score for simplicity.

For this reason, I will use F-score as the metric that considers both precision and recall.

Afonja, T. (2017). Accuracy Paradox. Towards Data Science, Retrieved December 3, 2018, from https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b

Project Design

Data exploration

I will first load the necessary Python libraries and load the marketing campaign data. The last column from this dataset, 'y', will be my target label (whether an individual

subscribes to the term deposit (yes/no). All other columns are features about each individual in the marketing campaign database.

A cursory investigation of the dataset will determine how many individuals fit into either group, and will tell us about the percentage of these individuals that purchased the product in a previous campaign.

Data preparation

Before data can be used as input for machine learning algorithms, it often must be cleaned, formatted, and restructured — this is typically known as preprocessing. Fortunately, for this dataset, there are no invalid or missing entries we must deal with, however, there are some qualities about certain features that must be adjusted. This preprocessing can help tremendously with the outcome and predictive power of nearly all learning algorithms.

Transforming Skewed Continuous Features

A dataset may sometimes contain at least one feature whose values tend to lie near a single number, but will also have a non-trivial number of vastly larger or smaller values than that single number. Algorithms can be sensitive to such distributions of values and can underperform if the range is not properly normalized. With the marketing campaign dataset two features fit this description: ‘age' and 'campaign'.

For highly-skewed feature distributions such as 'age' and 'campaign', it is common practice to apply a logarithmic transformation on the data so that the very large and very small values do not negatively affect the performance of a learning algorithm. Using a logarithmic transformation significantly reduces the range of values caused by outliers. Care must be taken when applying this transformation however: The logarithm of 0 is undefined, so we must translate the values by a small amount above 0 to apply the logarithm successfully.

Scaling from numerical variables

In addition to performing transformations on features that are highly skewed, it is often good practice to perform some type of scaling on numerical features. Applying a scaling to the data does not change the shape of each feature's distribution; however, normalization ensures that each feature is treated equally when applying supervised learners.

Encoding categorical variables

Learning algorithms expect input to be numeric, which requires that non-numeric features or categorical variables be converted. I will use one-hot encoding to create "dummy" variables for each possible category of each non-numeric feature.

Also, as with the non-numeric features, I will convert the non-numeric target label, 'label' to numerical values for the learning algorithm to work. Since there are only two possible categories for this label ("yes" and "no"), I don't need to use one-hot encoding and can instead simply encode these two categories as 0 and 1, respectively.

Data shuffle and split

Now all categorical variables have been converted into numerical features, and all numerical features have been normalized. I will now split the data (both features and their labels) into training and test sets. 80% of the data will be used for training and 20% for testing.

Six model comparison

Evaluating Model Performance

In this section, I will investigate six different algorithms with their default hyper parameters, and determine which is best at modeling the data.

Those algorithms are:

- Gaussian Naive Bayes (GaussianNB)
- Ensemble Methods (AdaBoost and Gradient Boosting)
- K-Nearest Neighbors (KNeighbors)
- Support Vector Machines (SVM)
- from sklearn.neural_network import MLPClassifier

Best model optimization

I will choose from the six supervised learning models the best model to use on the marketing campaign data. Then I will perform a grid search optimization for the model over the entire training set (X_train and y_train) by tuning several parameter to improve upon the untuned model's accuracy and F-score.

Feature selection

An important task when performing supervised learning on a dataset like the marketing campaign data, is determining which features provide the most predictive power. By focusing on the relationship between only a few crucial features and the target label I

can simplify my understanding of the phenomenon. For this reason, I will identify a small number of features that most strongly predict whether a client will subscribe to a term deposit.