Machine Learning Advanced Nanodegree

Capstone Project Proposal

Kailas Kasar

29th May 2018

- **Domain Background:**

Banking today has become one of the major industries, covering almost all aspects of life that we can think of. The various services and products made available are as per requirement of common customers. Bank reaches out to customers primarily through different channels like call centers, direct mail and face-to-face. Though Internet and Mobile marketing is now becoming common marketing tools for banks direct marketing is main strategy of many banks to interact with their customers.

Increase in digitalization has lead generation of gigabytes of data. Finding accurate insight for Marketing campaign using huge amount of data is becoming important for Banks.

It is very easy for Fortune 500 companies to plan as many as 3000 campaigns in a single year but still the effort are insignificant if they are not reaching prospects likely to increase deposit in bank. They can't afford to send direct mails to huge, undifferentiated databases. The frequency and turnaround of campaigns is higher than ever, and so is the expectation for return on investment.

The data mining has been used widely in direct marketing to identify prospective customers for new products, by using purchasing data, a predictive model to measure that a customer is going to respond to the promotion or an offer. Data mining has gained popularity for illustrative and predictive applications in banking processes.[1]

- **Problem Statement:**

Marketing campaigns are designed based on insights and decision derived from analyzing data. Analyzing huge amount of data using manual human efforts is nearly impossible. Machine Learning techniques like Logistic Regression, Decision Trees, and Random Forest become handy for predicting customers that have a higher probability to subscribe term deposit service offered by bank.

We will evaluate the performance of the classification models using statistical metrics like accuracy, sensitivity, precision etc. Model with better prediction with good accuracy is chosen to target the prospective customers.

- **Datasets and Inputs [2]**

Abstract: The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y)

Source:

[Moro et al., 2014] 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

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:
1. contact: contact communication type (categorical: 'cellular','telephone')
2. month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
3. day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
4. 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:
1. campaign: number of contacts performed during this campaign and for this client
(numeric, includes last contact)
2. 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)
3. previous: number of contacts performed before this campaign and for this client
(numeric)
4. poutcome: outcome of the previous marketing campaign (categorical:
'failure','nonexistent','success')
Social and economic context attributes
1. emp.var.rate: employment variation rate - quarterly indicator (numeric)
2. cons.price.idx: consumer price index - monthly indicator (numeric)
3. cons.conf.idx: consumer confidence index - monthly indicator (numeric)
4. euribor3m: euribor 3 month rate - daily indicator (numeric)
5. nr.employed: number of employees - quarterly indicator (numeric)

Output variable (desired target):
1. y - has the client subscribed a term deposit? (binary: 'yes','no')
Missing Attribute Values: There are several missing values in some categorical attributes, all coded with the "unknown" label. These missing values can be treated as a possible class label or using deletion or imputation techniques.

- **Solution Statement :**

Solution for Machine Learning Model to predict if a client will subscribe to the product term deposit will follow below given steps:

- **Exploring the Data :** Data exploration is need to check quality of data and to summarize their main characteristics, often with visual methods. Exploration of data is preliminary for seeing what the data can tell us beyond the formal modeling.
- **Data preprocessing :** Prepare the data by creating dummy variables for each of the categorical columns (since we cannot use textual data to build our model).
  e.g. housing. these can be reasonably converted into 1/0 (binary) values and other categorical columns like job with more than two values can assign with dummy variables and assign a 1 to one of them and 0 to all others.
- **Divide the data into a training set and a test set :** We will split the given training data in two sets ,70% of which will be used to train our models and 30% we will hold back as a test set.
- **Feature Scaling :[3]** Feature scaling can vary your results a lot while using certain algorithms and have a minimal or no effect in others. Most of the machine learning algorithms use Euclidian distance between two data points in their computations, this is a problem. Some examples of algorithms where feature scaling matters are: 1) K nearest neighbor 2) Naïve bayes 3) Tress based model.
- **Evaluate Algorithms:** Different machine learning techniques are available to solve classification problem. We will use 1) Logistic Regression (LR) 2) Classification and Regression Trees (CART) 3) Random Forests (RF) 4) Adaptive Boosting (AB)
  5) Extreme Gradient Boosting (XGB).
  We will use 10-fold cross validation to estimate accuracy. This will split our dataset 10 parts, train on 9 and test on 1 and repeat for all combinations of train-test splits.
  We will use the metric of accuracy to evaluate models. This is a ratio of the number of correctly predicted instances in divided by the total number of instances in the dataset

- **Benchmark Model :**

The given dataset is a typical supervised learning problem for which tree type models perform a lot better than the rest. So we will pick Extreme Gradient Boosting (XGB) as benchmark and try to beat the benchmark with hyperparameter turning.

- **Evaluation Metrics :**

Different performance metrics are used to evaluate different Machine Learning Algorithms. Choice of metrics influences how the performance of machine learning algorithms is measured and compared.

Confustion Matrix :

Confusion Matrix or Error matrix : It is a table that describes the performance of a supervised machine learning model on the testing data, where the true values are unknown. Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (and vice versa).

Let us consider the following confusion matrix

a is the number of correct predictions that an instance is negative,

b is the number of incorrect predictions that an instance is positive,

c is the number of incorrect of predictions that an instance negative, and

d is the number of correct predictions that an instance is positive.

		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

The **accuracy** (AC) is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$AC = \frac{a+d}{a+b+c+d}$$

The **recall** or true positive rate (TP) is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$TP = \frac{d}{c+d}$$

The **false positive** rate (FP) is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation:

$$FP = \frac{b}{a+b}$$

The **true negative** rate (TN) is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation:

$$TN = \frac{a}{a+b}$$

The **false negative** rate (FN) is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation:

$$FN = \frac{c}{c+d}$$

Finally, **precision (P)** is the proportion of the predicted positive cases that were correct, as calculated using the equation:

$$P = \frac{d}{b+d}$$

For this classification problem, I will use confusion matrix, accuracy, precision and recall to evaluate the performance of the machine learning models.

- **Project Design**

The workflow of solving this problem will be in the following order:

- Exploring the Data
  - Loading Libraries and data
  - Peek at the training data
  - Dimensions of data
  - Overview of responses and overall response rate
  - Statistical summary
- Data preprocessing
  - Preprocess feature columns
  - Identify Feature and Target columns
  - Data cleaning
  - Training and Validation data split
  - Feature Scaling - Standardization/Normalizing data
- Evaluate Algorithms
  - Build models
  - Select best model
  - Make predictions on the validation set
  - Feature importance and feature selection
- Model Tuning to Improve Result
- Final conclusion

References :

[1] https://analyticsindiamag.com/analytics-in-indian-banking-sector-on-a-right-track/

[2] https://archive.ics.uci.edu/ml/datasets/bank+marketing

[3] https://en.wikipedia.org/wiki/Feature_scaling

[4] https://medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e