Machine Learning Engineer Nanodegree

Capstone Proposal

Malgorzata Kot
June 26th, 2018

Proposal

Domain Background

Marketing used to be a creative discipline, which enables companies to stand out after campaigns based on creativity of marketeers. Since then, it was changed moslty with the development of technology. There are already multiple companies taking benefit of machine learning algorithms, which support or even replace marketeers. One of clear examples, where company believes in algorithms much more than in creative marketing, is Zalando, which replaces 200 marketing employes with AI and machine learning, which will be implemented by data scientists.[1] According to research, Zalando's decision was reasonable - sudies show, that using machine learning models for customer classification results in higher conversion rates[2], better potential customer identification[3] or can be used for quick sentiment classification [4], which was previosuly done manually. One of the problems, marketing is facing every day, is predicting, who of the potential customers will buy a product. It helps in efficient allocation of ressources as well as getting to know the customer base. For companies with large numbers of customers with heterogenous characteristics, supervised machine learning can be used for predicting potential customer interest in buying a product.

Supervised machine learning, also called classification, uses past outcome for predicting future. It means that for every training (and test) example, there is a real life result, which for marketing can be if the client has purchased or not.[5] Bassed on this information, clients can be classified as a merketing opportunity and audience without a buying interest. It leads to better identifying potential buyers, which spares time as only people with potential interest in buying the product will be contacted. There are numerours machine learning algorithms, which can be used for classification problems, as boosting, random forest, bagging or SVMs and most of them shows also very good performances over different metrics (accuracy, F-Score and other)[6].

Problem Statement

In this project, I will classify potential customers of a bank based on an information if they bought a product

or not. The classifier after getting data from a potential customer should return a 1 or 0 depending if the customer is likely to buy a product and a probabilty based on which the class was computed. Classification algorithm will be measured based on accuracy, precision, recall and F-Beta Score, which should be higher than in the base model.

Datasets and Inputs

This project will use Bank Marketing Dataset from Machine Learning Repository (link). This dataset represent marketing activities from a bank.

The dataset contains of 20 input variables and out output variable. Variables' descriptions were taken from Machine Learning Repository Website, linked above.

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

Output variable:
21 - y - has the client subscribed a term deposit? (binary: ‘yes’,’no’).

Output variable gives an information if the person bought a product (subscribed a term deposit in this dataset). Input variables are actually a description of a person in data, description of campaign, contact time and additional attributes. These attributes are most likely the data which most businesses own and use when advertising their products.

Solution Statement

As mentioned briefly in domain background, this problem will be solved by using supervised classification algorithms. I will test a few algorithms out of: naive bayes, decision trees, SVMs, logistic regression, bagged algorithm as random forest, boosted algorithm as adaboost and neural network. I will choose the algorithm which performs best in accuracy and F-Beta score. Further more, I will make an analysis of the most important features and try to reduce the amount of features needed for final prediction.

Benchmark Model

Benchmark model is based on a given data. I assume, that marketeers saw potential in every potential customer they called, so to measure if model will be more accurate than this, I will count how many per cent of called people actually bought a product.

benchmark accuracy = people who bought the product / all people called

In the same way I will compute F-Beta score for existing data, where true positives are poeple who did buy a product, there will be no false positives, true negatives are people called who did not buy the product and also, there are no false negatives.

Evaluation Metrics

To evaluate the model, I will use accuracy and F-Beta score.

Accuracy is the amount of correctly predicted values divided by all predictions and multiplied by 100. Accuracy is not the most useful measure for this problem. If the assumption in benchmarm model is, that all contacted clients are willing to buy, the model would get a low accuracy (11%). But if the assumption would be, that none of called people would by a product, accuracy will be very high (89%). Where second predictions has a very high accuracy, it makes completely no sense, as we look for people who are willing to buy a product. First prediction makes more sense, but the number is very low and easy to improve. So if final model would have 15% accuracy it would be already 30% better thank benchmark but still bad. The problem is, that it is very hard to say where bad accuracy ends and good accuracy starts.

In order to get to know the results of the model a bit more, I will use precision and recall combined in F-Beta score. Precision measures how many true prediction were made out all values predictied true (true positives and false positives). Recall measures how many true positives were out of all true values in dataset. F-Beta score combines both precision and recall in order to find the balance between them.

$$(10.1)\quad Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$(10.2)\quad Precision = \frac{T_p}{T_p + F_p}$$

$$(10.3)\quad Recall = \frac{T_p}{T_p + T_n}$$

$$(10.4)\quad F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Fig. 1: Accuracy, Precision, Recall and F-Beta score[7]

Because this is marketing data, where we look for potential opportunities, identifying all potential customers is more improtant than making sure, that they are no false identifications. That is why I will use F-2 score, where beta = 2, what give more credit to recall than precision.

Additionally, I will create confusion matrix in order to see how many false/true positives and negatives are predicted.

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

Fig. 2: Confusion matrix[8]

Project Design

Data preparation

The dataset has to be prepared for analysis. That means, output variable has to be boolean. If there are any other text variables which can be translated to numbers, it will be also done, otherwise variables will be translated into boolean columns using dummies.

Exploratory data analysis

Exploratory data analysis will be second step in order to check features distributions and if some features are highly correlated and could be removed from dataset. There will also be performed outlier detection and checking if there are any features correlating with each other. It will include second part of the data preparation - outlier removal, log transformations for skewed distributions, removing or creating new features based on the dataset (all if needed). As features in the dataset have different ranges, most likely there will be also some type of feature scaling necessary (min-max normalization or similar.)

Model selection

I plan to test at least three different models as SVMs, decision trees and logistic regression or other. I will use shuffle split for splitting data into test and training. Because this dataset is quiet large (40k data points), I may do a few different shuffle splits, but it is not mandatory. In order to choose the best algorithm, I will also use

grid search cv, which will help me choosing best hyperparameters.

Model evaluation

I will choose the model with highest F-Beta score, but I will also consider accuracy, confusion matrix and model predicting time.

Feature selection

If model will alow it, I will extract feature importances and try to reduce number of features which model will be using possibly without loss in evaluation metrics.

---

[1] https://www.thelocal.de/20180309/zalando-to-replace-up-to-250-jobs-with-algorithms

[2] Pål Sundsøy, Johannes Bjelland, Asif M Iqbal, Alex Sandy Pentland, Yves-Alexandre de Montjoye. *Big Data-Driven Marketing: How machine learning outperforms marketers' gut-feeling.* In *Social Computing, Behavioral-Cultural Modeling & Prediction Lecture Notes* in *Computer Science* Volume 8393. 2014, pp. 367-374.

[3] Dirk Thorleuchter, Dirk Van den Poel, Anita Prinzie. *Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing.* In *Expert Systems with Applications* Volume 39(3). February 2012, pp. 2597-2605.

[4] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. *Thumbs up? Sentiment Classification using Machine Learning Techniques* In *Proceedings of EMNLP*. 2002, pp. 79–86.

[5] Ian H. Witten, Eibe Frank, Mark A. Hall, Christopher J. Pal. *Data Mining: Practical Machine Learning Tools and Techniques, Fourth Edition.* Morgan Kaufmann, 1 Oct 2016, pp. 45

[6] Rich Caruana, Alexandru Niculescu-Mizil. *An Empirical Comparison of Supervised Learning Algorithms* In *Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, PA, 2006.

[7] https://www.safaribooksonline.com/library/view/python-data-analysis/9781785282287/graphics/B04223_10_02.jpg

[8] http://3.bp.blogspot.com/_txFWHHNYMJQ/THyADzbutYI/AAAAAAAAAf8/TAXL7lySrko/s1600/Picture+8.png