Machine Learning Engineer Nanodegree

Capstone Project

Malgorzata Kot
29.07.2018

I. Definition

Project Overview

Machine Learning can support diverse discplines and one of them is marketing. For example, according to research using machine learning models for customer classification results in higher conversion rates[1], better potential customer identification[2] or can be used for quick sentiment classification [3], 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.

In this project, potential customers will be classified based on bank marketing dataset from UCI Machine Learning Repository. (link) The goal is to predict the variable “y”, which explains if customer has subscribed a deposit or not, which is equivalent to: did he buy a product or not.

Problem Statement

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, 1 meaning customer will buy the product and 0 he/she will not buy. Additionally, there will be probabilty returned of how likely it is, that customer will subscribe to the deposit.

Metrics

Main metric will be F-2 Score (derived from F-Beta score), but additionally it will be measured with accuracy and confusion matrix. Ideally all of them should be higher than in the base model.

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. 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.

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 benchmark 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.

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

II. Analysis

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.

Data Exploration

Dataset contains 21 variables, of which 1 is output variable and 41188 observations.

```
In [28]: dataset_df.shape
Out[28]: (41188, 21)
```

Fig. 1: Dataset shape

Out of 20 input variables in the dataset half is numerical and another one is categorical, stored as objects (strings).

```
In [26]: dataset_df.dtypes

Out[26]: age               int64
         job              object
         marital          object
         education        object
         default          object
         housing          object
         loan             object
         contact          object
         month            object
         day_of_week      object
         duration          int64
         campaign          int64
         pdays             int64
         previous          int64
         poutcome         object
         emp.var.rate    float64
         cons.price.idx  float64
         cons.conf.idx   float64
         euribor3m       float64
         nr.employed     float64
         y                object
         dtype: object
```

Fig. 2: Data types in dataset

```
dataset_df.head()
```

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	261	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
1	57	services	married	high.school	unknown	no	no	telephone	may	mon	149	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
2	37	services	married	high.school	no	yes	no	telephone	may	mon	226	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	151	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no
4	56	services	married	high.school	no	no	yes	telephone	may	mon	307	1	999	0	nonexistent	1.1	93.994	-36.4	4.857	5191.0	no

Fig. 3: Datset sample

Out of numerical variables, only cons.conf.idx and cons.price.idx look like they would not be skewed. There are also big scale differences between them. Eg. ‘emp.var.rate’ varies from -3.4 to 1.4 and duration from 0 to 4918.

Most of the categorical variables habe 2 to 4 options, apart of job and education as well as date related variables. Some of categorical variables seem to have not categorized data, eg. ‘poutcome’ features has ‘nonexistent’ as the most often input.

```
dataset_df.describe(include='all')
```

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed	y
count	41188.00000	41188	41188	41188	41188	41188	41188	41188	41188	41188	41188.000000	41188.000000	41188.000000	41188.000000	41188	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	41188
unique	NaN	12	4	8	3	3	3	2	10	5	NaN	NaN	NaN	NaN	3	NaN	NaN	NaN	NaN	NaN	2
top	NaN	admin.	married	university.degree	no	yes	no	cellular	may	thu	NaN	NaN	NaN	NaN	nonexistent	NaN	NaN	NaN	NaN	NaN	no
freq	NaN	10422	24928	12168	32588	21576	33950	26144	13769	8623	NaN	NaN	NaN	NaN	35563	NaN	NaN	NaN	NaN	NaN	36548
mean	40.02406	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	258.285010	2.567593	962.475454	0.172963	NaN	0.081886	93.575664	-40.502600	3.621291	5167.035911	NaN
std	10.42125	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	259.279249	2.770014	186.910907	0.494901	NaN	1.570960	0.578840	4.628198	1.734447	72.251528	NaN
min	17.00000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.000000	1.000000	0.000000	0.000000	NaN	-3.400000	92.201000	-50.800000	0.634000	4963.600000	NaN
25%	32.00000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	102.000000	1.000000	999.000000	0.000000	NaN	-1.800000	93.075000	-42.700000	1.344000	5099.100000	NaN
50%	38.00000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	180.000000	2.000000	999.000000	0.000000	NaN	1.100000	93.749000	-41.800000	4.857000	5191.000000	NaN
75%	47.00000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	319.000000	3.000000	999.000000	0.000000	NaN	1.400000	93.994000	-36.400000	4.961000	5228.100000	NaN
max	98.00000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4918.000000	56.000000	999.000000	7.000000	NaN	1.400000	94.767000	-26.900000	5.045000	5228.100000	NaN

Fig. 4: Dataset variables description

Exploratory Visualization

In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should adequately support the data being used. Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section:

- *Have you visualized a relevant characteristic or feature about the dataset or input data?*
- *Is the visualization thoroughly analyzed and discussed?*
- *If a plot is provided, are the axes, title, and datum clearly defined?*

Algorithms and Techniques

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

- *Are the algorithms you will use, including any default variables/parameters in the project clearly defined?*
- *Are the techniques to be used thoroughly discussed and justified?*
- *Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?*

Benchmark

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section:

- *Has some result or value been provided that acts as a benchmark for measuring performance?*
- *Is it clear how this result or value was obtained (whether by data or by hypothesis)?*

III. Methodology

(approx. 3-5 pages)

Data Preprocessing

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:

- *If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?*
- *Based on the **Data Exploration** section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?*
- *If no preprocessing is needed, has it been made clear why?*

Implementation

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

- *Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?*
- *Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?*
- *Was there any part of the coding process (e.g., writing complicated functions) that should be documented?*

Refinement

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

- *Has an initial solution been found and clearly reported?*
- *Is the process of improvement clearly documented, such as what techniques were used?*
- *Are intermediate and final solutions clearly reported as the process is improved?*

IV. Results

(approx. 2-3 pages)

Model Evaluation and Validation

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be

used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model's solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section:

- *Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?*
- *Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data?*
- *Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results?*
- *Can results found from the model be trusted?*

Justification

In this section, your model's final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section:

- *Are the final results found stronger than the benchmark result reported earlier?*
- *Have you thoroughly analyzed and discussed the final solution?*
- *Is the final solution significant enough to have solved the problem?*

V. Conclusion

(approx. 1-2 pages)

Free-Form Visualization

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section:

- *Have you visualized a relevant or important quality about the problem, dataset, input data, or results?*
- *Is the visualization thoroughly analyzed and discussed?*
- *If a plot is provided, are the axes, title, and datum clearly defined?*

Reflection

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole

to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section:

- *Have you thoroughly summarized the entire process you used for this project?*
- *Were there any interesting aspects of the project?*
- *Were there any difficult aspects of the project?*
- *Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?*

Improvement

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section:

- *Are there further improvements that could be made on the algorithms or techniques you used in this project?*
- *Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how?*
- *If you used your final solution as the new benchmark, do you think an even better solution exists?*

Before submitting, ask yourself. . .

- Does the project report you've written follow a well-organized structure similar to that of the project template?
- Is each section (particularly **Analysis** and **Methodology**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
- Would the intended audience of your project be able to understand your analysis, methods, and results?
- Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?
- Are all the resources used for this project correctly cited and referenced?
- Is the code that implements your solution easily readable and properly commented?
- Does the code execute without error and produce results similar to those reported?

[1] 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.

[2] 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.

[3] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. *Thumbs up? Sentiment Classification using Machine*

Learning Techniques In *Proceedings of EMNLP*. 2002, pp. 79–86.