Data Science Take-Home Exercise

Clear Brain

Scooter Ride Share Service

Elias BOUGRINE

05/30/2019

Abstract

In this project, the task is to analyze one dataset in order to build a prediction model that will predict if a user is going to buy a service or not.
Our model will thus return 1 if the probability than a user will be a customer is superior than the probability that the user won't buy the service. The model will return 0 in the opposite case.

Methods

We will follow the following steps:

- Exploratory Data Analysis
- Data Cleaning and Feature Preprocessing
- Classification
- Evaluating Classifiers
- Testing the Model on the Validation Dataset
- Conclusions

1. Exploratory Data Analysis

The Exploratory Data Analysis is a major component of the project since it permits us to summarize, visualize and transform data in order to understand them more deeply.

Here are some of the conclusions I made after processing this step.

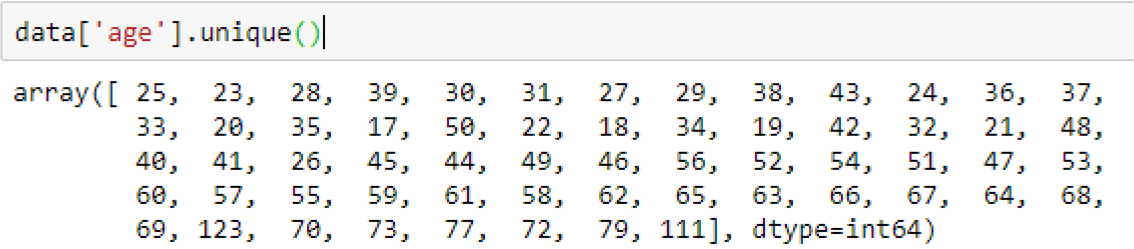
a. Country

The users come from four different countries: UK, US, China or Germany.

b. Age

The analysis of the different ages of the users showed that some users are 111 or 123 years old. It seems to be an error due to the acquisition of the data.

Otherwise, the ages of the users vary from 25 to 79 years old.



```
data['age'].unique()
```

```
array([ 25,  23,  28,  39,  30,  31,  27,  29,  38,  43,  24,  36,  37,
        33,  20,  35,  17,  50,  22,  18,  34,  19,  42,  32,  21,  48,
        40,  41,  26,  45,  44,  49,  46,  56,  52,  54,  51,  47,  53,
        60,  57,  55,  59,  61,  58,  62,  65,  63,  66,  67,  64,  68,
        69, 123,  70,  73,  77,  72,  79, 111], dtype=int64)
```

c. Total Pages Visited

The total number of pages visited by a user during the session varies from 1 to 29.

d. New User/ Source/ Converted

The data in these three columns were coherent and in the same formats than the ones specified in the README file.

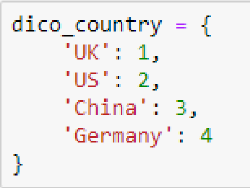
2. Data Cleaning and Feature Preprocessing

Data come in many formats and vary greatly in usefulness for analysis. The term ‘data cleaning’ refers to the process of combing through the data and deciding how to resolve inconsistencies and missing values for example.

In order to get clean features to build a prediction model, I processed a data cleaning of the dataset:

a. Country

The country column initially contained a string value. In order to use these data as a feature, I changed these string values into integers:



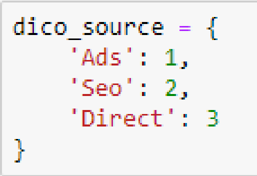
```
dico_country = {
    'UK': 1,
    'US': 2,
    'China': 3,
    'Germany': 4
}
```

b. Age

We noticed during the Exploratory Data Analysis that some customers are 111 or 123 years old. This seems to be an error. In order to have a better accuracy when building our prediction model, we will not take these rows into consideration.

c. Source

We want to have integer values instead of string values. We affect each of the unique source to an integer through the mean of a dictionary and replace in the Data Frame the string by these integers.



```
dico_source = {
    'Ads': 1,
    'Seo': 2,
    'Direct': 3
}
```

3. Classification

Classification is the process of making categorical predictions based on data. We may reconstruct classification as a type of regression problem. Instead of creating a model to predict an arbitrary number, we create a model to predict a probability that a data belongs to a category. This allows us to reuse the machinery of linear regression for a regression on probabilities: logistic regression.

In order to create a model and evaluate its accuracy on a validation set, the first step is to split the dataset into a training and validation sets.

a. Training Validation Split

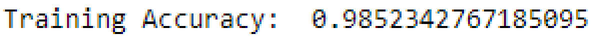
The data we have is all the data we have available for both training the model and validating the model that we train. We therefore need to split the data into separate training and validation datasets. We will need this validation data to assess the performance of our classifier once we are finished training.

b. Retrieve the features matrix and the model output vector

In order to fit a model, we need to create a matrix containing the features and a vector containing the real values corresponding to the output values we have knowing the features.

c. Fit the model

We fit a Logistic Regression model with these vectors, and we look at the training accuracy.



```
Training Accuracy:  0.9852342767185095
```

I obtained an accuracy of 98.526%. Even if it seems very accurate, it is not a good way to evaluate the precision of the classifier.

4. Evaluating the Classifiers

First, we are evaluating accuracy on the training set, which may lead to a misleading accuracy measure, especially if we used the training set to identify discriminative features.

Presumably, our classifier will be used for filtering. There are two kinds of errors we can make:

- False positive (FP): A user that will not buy the service and that will be considered as converted.
- False negative (FN): A user that will be buy the service and that won't be considered as converted.

These definitions depend both on the true labels and the predicted labels. False positives and false negatives may be of differing importance, leading us to consider more ways of evaluating a classifier, in addition to overall accuracy:

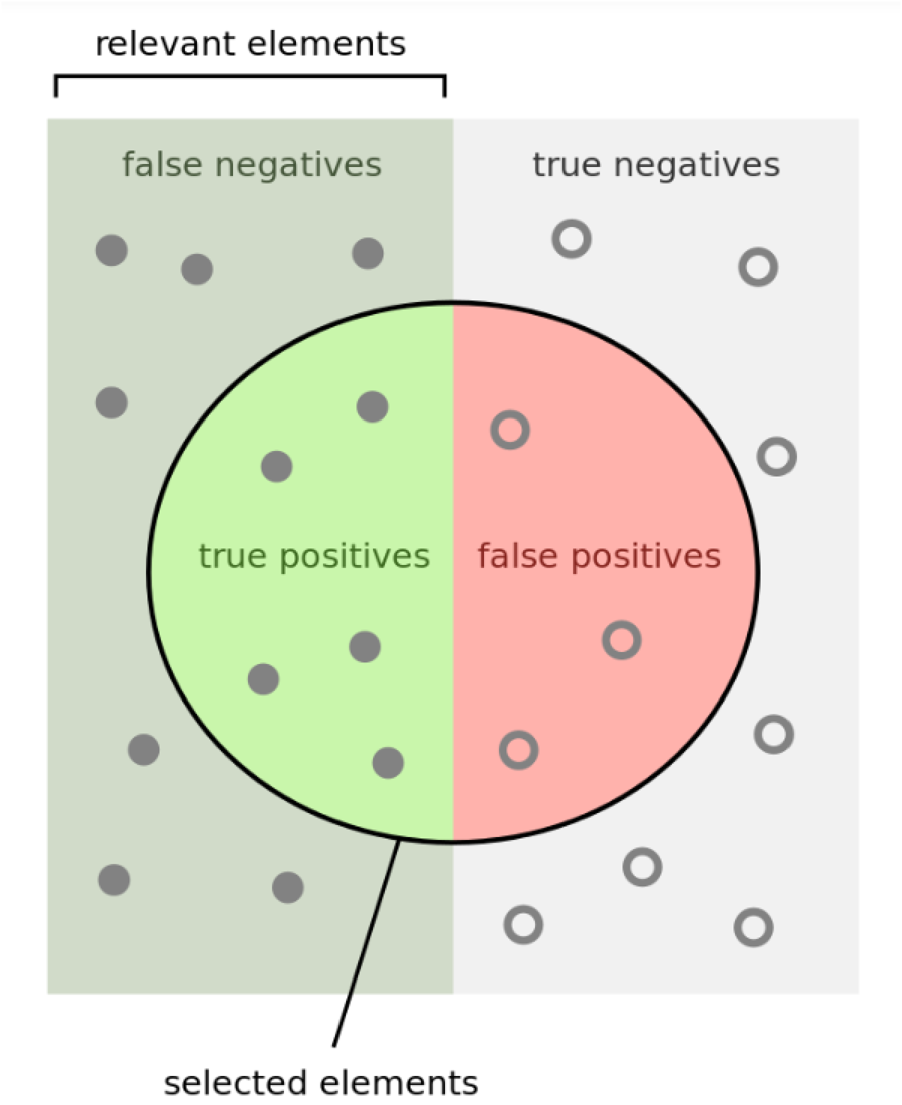
- **Precision**: measures the proportion of predicted conversions that are actually conversions.

- **Recall**: measures the proportion of conversions that were correctly flagged as conversions.

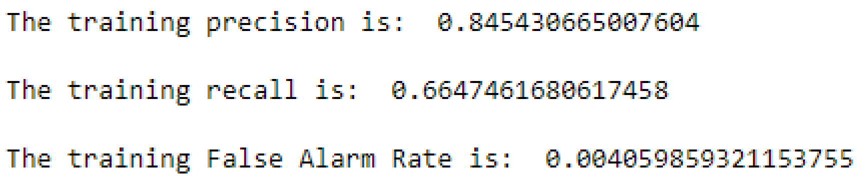
- **False alarm rate**: measures the proportion of non-conversions that were incorrectly flagged as conversions.

Note that a true positive (TP) is a user that will buy the service and will be considered as converted, and a true negative (TN) is a user that won't buy anything and that won't be considered as converted.

The following image summarizes these errors:



Knowing these definitions, we have the following results:



```
The training precision is:  0.845430665007604

The training recall is:  0.6647461680617458

The training False Alarm Rate is:  0.004059859321153755
```

The training precision of our classifier is now 84.54%.

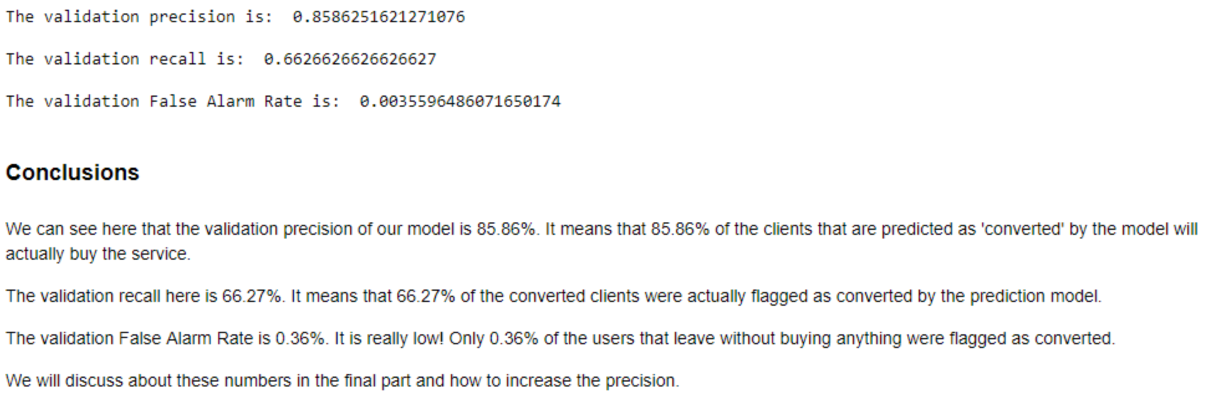
5. Testing the Model on the Validation Dataset

Now, we test the model on the validation dataset to obtain the validation precision.

We therefore did the same steps:

- Retrieve X_val and Y_val
- Use the Model we built with the training dataset
- Evaluate the classifier

We obtain the following results:



```
The validation precision is:  0.8586251621271076

The validation recall is:  0.6626626626626627

The validation False Alarm Rate is:  0.0035596486071650174
```

Conclusions

We can see here that the validation precision of our model is 85.86%. It means that 85.86% of the clients that are predicted as 'converted' by the model will actually buy the service.

The validation recall here is 66.27%. It means that 66.27% of the converted clients were actually flagged as converted by the prediction model.

The validation False Alarm Rate is 0.36%. It is really low! Only 0.36% of the users that leave without buying anything were flagged as converted.

We will discuss about these numbers in the final part and how to increase the precision.

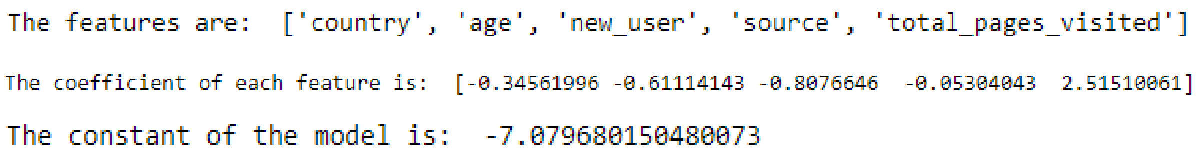
6. Conclusions

a. Accuracy and precision

As we said before, even if the accuracy seems really high (98.59%), the real precision of the classifier is only 85.86%. The precision could be improved by adding more features.

b. Coefficients of the model and importance of the features

We have the following coefficients:



```
The features are:  ['country', 'age', 'new_user', 'source', 'total_pages_visited']

The coefficient of each feature is:  [-0.34561996 -0.61114143 -0.8076646  -0.05304043  2.51510061]

The constant of the model is:  -7.079680150480073
```

Since we normalized the features, we can analyze and compare their coefficients:

- Here, we can see that the total number of visited pages has a big importance on deciding if a user is going to buy the service or not.
- If a user is new to the platform, he is not willing to buy the service.
- The age of the user is also an important feature: the older the user is, the less he is going to buy the service.
- Finally, the country of the user and the source he is using are not important features.

c. Recommendations and improvements

In order to improve the model, it could be useful to have more features such as:

- The gender of the user
- The price of the service
- The total duration of the ride

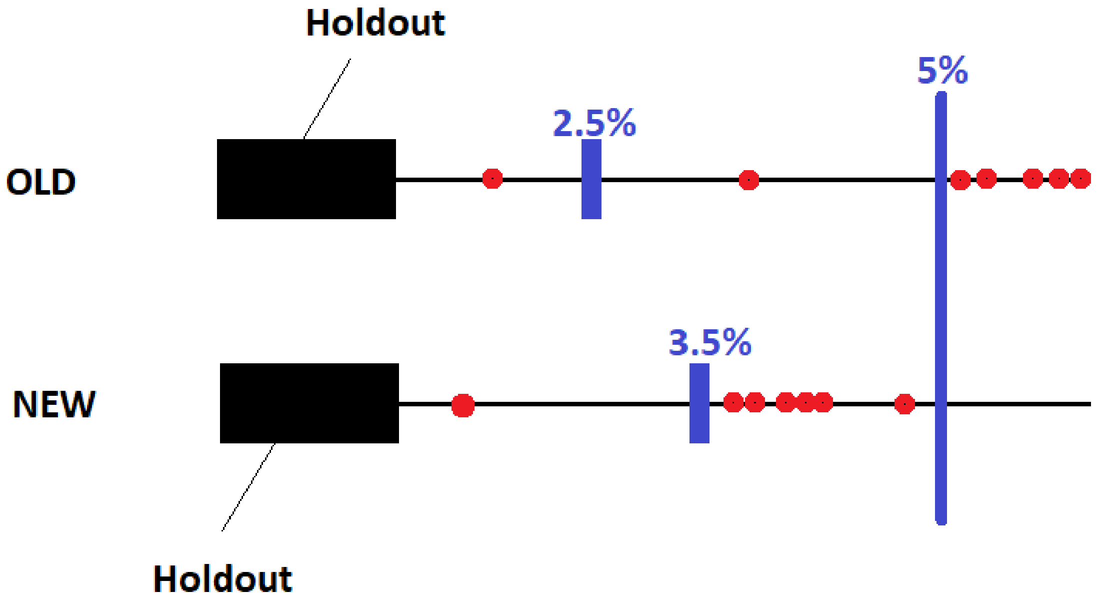
These features could improve the precision of the classifier.

New Marketing Methodology

In this case, we cannot conclude whether the new marketing methodology will provide an increase in incremental conversions over the old methodology.

For both the new and old marketing methodology, we have a statistical significance of 99%. This means for instance that the incremental conversion rate above its holdout for the old technology will be superior or equal to 2.5% with a probability 0.99.

Let’s now take an example showing that we cannot conclude with only this information:



That example presents a case where the old methodology has an incremental conversion rate above its holdout of 2.5% with a statistical significance of 99%.

The new methodology has an incremental conversion rate above its holdout of 3.5% with a statistical significance of 99%.

However, the new methodology has statistically a better conversion rate.