Task for Quantitative Analyst

Tomas Uždavinys

December 21, 2021

Contents

1 Task 2 **1**
1.1 Observations 1
1.2 Visualizations 2

2 Task 3 **2**
2.1 Model evaluation metrics 2
2.2 Feature transformation 7
2.3 Baseline predictions 7
2.4 Logistic regression 8

1 Task 2

Visualize data. Free to select the scope, types of plots, etc.

1.1 Observations

List of qualitative observations from analyzing data visualizations.

- Features *pdays* and *previous* encode the same information, whatever customer was contacted before this campaign.
- It appears that younger and older people are more likely to subscribe to long-term deposit contracts.
- Having a larger balance (or at least above median) suggests that is directly proportional to subscriptions proportion.
- From visual data inspection, people with no personal and housing loans are more likely to subscribe. This group represents a substantial proportion of data (38%).

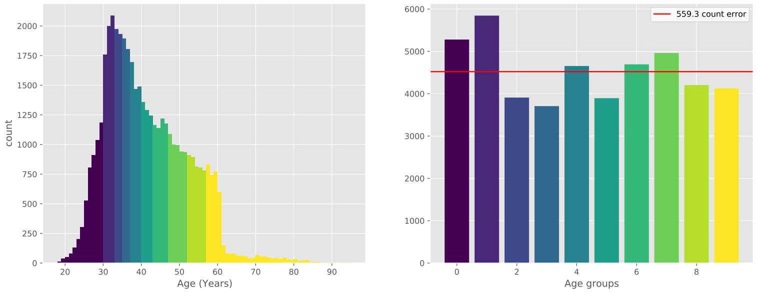

Figure 1: Age feature was cut into 10 deciles, almost equal size subsets. Age distributions: histogram (left) and quantiles barplot (right).

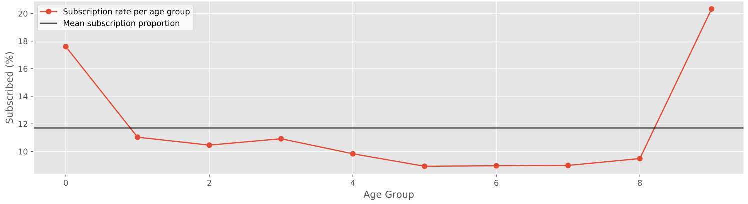

Figure 2: From visual inspection, it seems that younger and older customers are more likely to subscribe.

1.2 Visualizations

3 functions were created to visualize numerical features. Other visualization graphs were manually plotted using *matplotlib* or *seaborn* packages depending on question.

2 Task 3

Perform a logistic regression to obtain the predicted probability that a customer has subscribed for a term deposit. Use continuous variables and dummy variables created for categorical columns. Not necessarily all variables provided in data sample should be used. Evaluate model goodness of fit and predictive ability. If needed, data set could be split into training and test sets.

2.1 Model evaluation metrics

Before training a new model, one needs to create some baseline predictions and/or simple models. In addition, depending on business problem evaluation metrics should be established. It is not explicitly stated, what requirements are

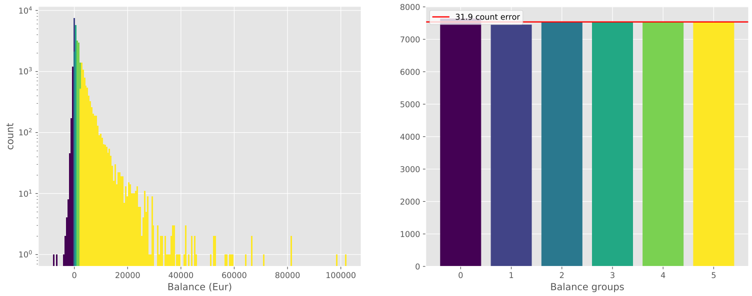

Figure 3: The balance feature (average yearly balance, in euros) was cut into 6 almost equal quantiles.

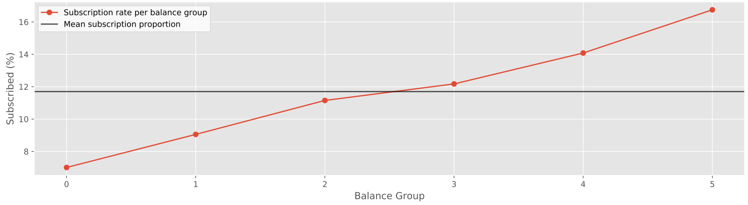

Figure 4: The likelihood of subscribing to long-term deposits increases with the account balance.

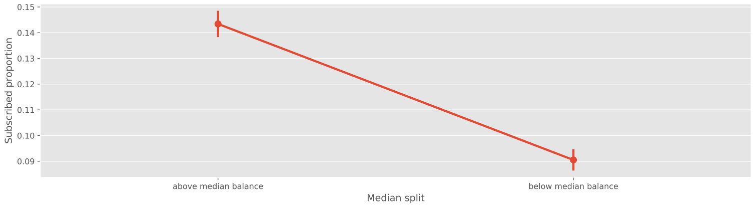

Figure 5: By splitting customers into 2 groups: balance above sample median and below, this trend becomes more pronounced.

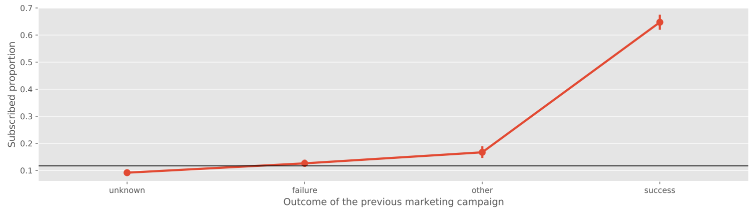

Figure 6: People who previously successfully subscribed to long-term deposits are much more likely to subscribe again.

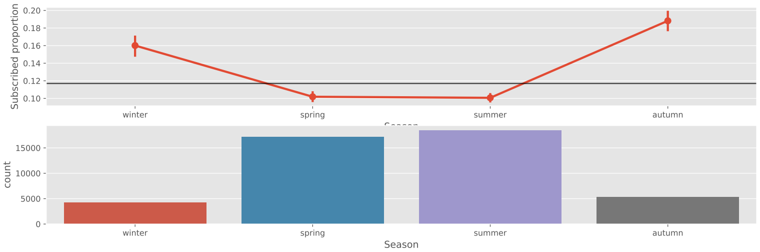

Figure 7: More people were contacted during spring and summer than winter and autumn months. While the subscription proportion is larger for later seasons, we can't directly compare these proportions because of imbalanced data. Statistical methods are required to evaluate whatever season has a statistically meaningful impact on the subscription rate increase.

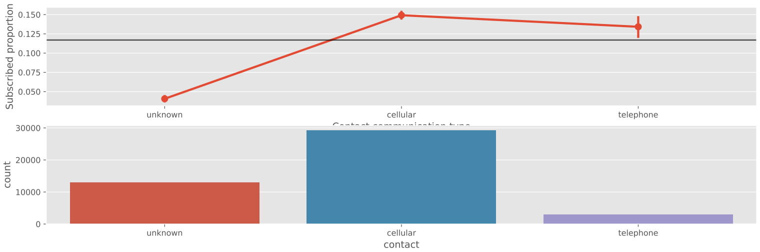

Figure 8: Contacting people via cellular or telephone also increases customer subscription proportion.

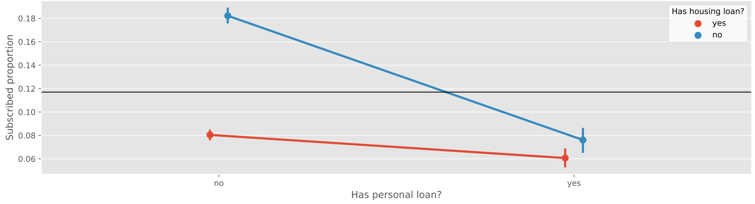

Figure 9: People with no loans are more likely to subscribe.

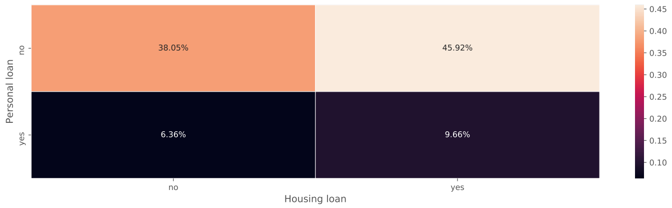

Figure 10: A substantially large proportion (38%) of customers doesn't have any loans.

for the logistic model. I will provide a quick overview of the most important metrics for classification problems and which ones to use. I also dropped *previous* feature as it encodes basically same information as *pdays*.

- **Accuracy-** measures how often the classifier makes the correct prediction. From the exploratory data analysis, we know that 11.7% of customers subscribed to long-term deposits., therefore by contacting all customers minimum accuracy of 11.7% can be achieved. Accuracy of 88.3% can be achieved by making all predictions equal to zero. Accuracy alone is not a sufficient metric to evaluate logistic regression model.

- **Precision-** tells us what proportion of customers were classified as 1 (subscribed to take a long-term deposit) actually subscribed. It is a ratio of true positives to all positives. For our blind predictions above, precision will be equal to the accuracy score. For new models, we would like to have a precision score larger than 11.7%.

- **Recall-** tells us what proportion of customers that actually subscribed were classified by a model as subscribers. A low recall score would mean, that a lot of potential customers were not contacted.

- **F1-** is used to combine the precision and recall metrics into a single metric. Increasing the model's precision usually leads to decreased recall, i.e. fewer true customers will be contacted, and vice versa. F-1 score calculates a harmonic mean of these 2 scores, i.e average performance of precision

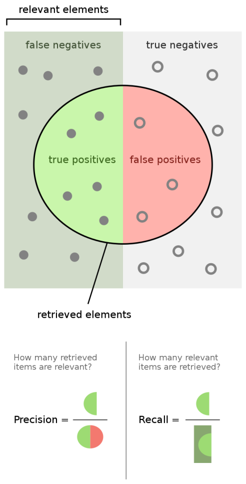

Figure 11: Precision and recall illustration.

and recall metrics. I will be using generic F1 expression, which equally prioritizes recall and precision.

- **Brier score-** measures the accuracy of probabilistic predictions. Call centers might have the capacity to call every client, but they want to prioritize calls based on the model's probability predictions, i.e. first call to customers with the highest probability of success. While I will not use the brier score in this assignment, it is a very useful metric for calibrating the model's over- and under-estimations. E.g. model provides 30% prediction that customers will subscribe. In reality, 40% of customers subscribed with such a prediction. Underestimated probabilities would lead to wrong client prioritizing.

Models will be evaluated using accuracy, precision, recall, and F-1 scores. Further optimizations should be conducted based on customer needs.

- If returns from successful subscriptions out weights the costs of subscribing a client recall and accuracy metrics should be used for evaluating the model.
- If our goal is to minimize the number of calls then precision should be used instead.
- For this task, I will focus on F-1 score as it allows me to address both needs expressed above.

2.2 Feature transformation

All data transformation is listed in function *transform_df*. More details can be found inside the function as comments. For first predictions, I try to avoid transformations on numerical data, e.g. applying log transform to normalized skewed features. Instead, I cut data into quantiles and applied min/max normalization. Categorical features were one-hot-encoded.

2.3 Baseline predictions

The first baseline prediction will be to predict all customers as subscribers. Such prediction would have the following results:

- Accuracy: 0.117,
- Precision: 0.117,
- Recall: 1.000,
- F-1: 0.209.

From visual data analysis (**TASK 2**) I identified 3 features (no statistical tests were conducted):

- customer belongs to 1st and 10th deciles, i.e. are younger than 29 and older than 56,
- customers who don't have any loans (personal and housing),
- customers with average yearly account balance above sample median ($>$ 448 euros).

73.2% were classified as potential subscribers. Such naive prediction resulted in the following scores:

- Accuracy: 0.352,
- Precision: 0.137,
- Recall: 0.859,
- F-1: 0.237.

Compared to blindly calling every customer, we managed to capture 85% of all potential clients by only contacting 73.2% of customers with more than double accuracy. Based on the increased F-1 score, precision increased more than recall decreased.

Then training ML models, logistic regression, or other, one must always split data into 2-subsets: train and test. All scores below are calculated on test data after training model on train dataset.

Before selecting features or choosing the best ML for the problem, I train a naive Bayes classifier on all transformed features. It is a quick and rather easy to interpret model to establish baseline prediction.

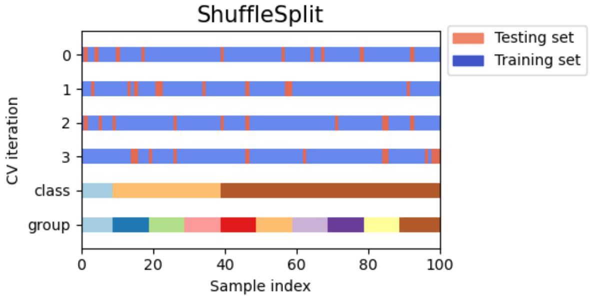

Figure 12: Shuffle split visualization. Data is N times randomly split into train and test subsets. Then the model is N times trained and evaluated. All data is used during each cross-validation iteration. Splitting is done randomly.

- Accuracy: 0.855,
- Precision: 0.472,
- Recall: 0.402,
- F-1: 0.434.

While recall drastically dropped, our precision increased to almost 50%. As we are trying to capture both metrics, F-1 score doubled compared to the most basic prediction, hence the naive Bayes model is better than our predictions based on visualization results.

2.4 Logistic regression

I used the cross-validation shuffling technique to evaluate features. I ran 5 iterations, at each, I cross-validated the model without 1 feature. The highest F-1 score was used to decide which feature should be dropped after each run. Cross-validation was conducted on train data set.

- After the first iteration, the average F-1 score without the day feature was 0.411. Day feature is dropped.
- After 2nd run, the education feature was dropped, the maximum average F-1 increased to 0.413.
- the default feature is dropped. The maximum average F-1 score is 0.414.
- the pdays feature is dropped. The maximum F-1 score is 0.414.
- the age feature is dropped. The maximum average F-1 score is 0.412.

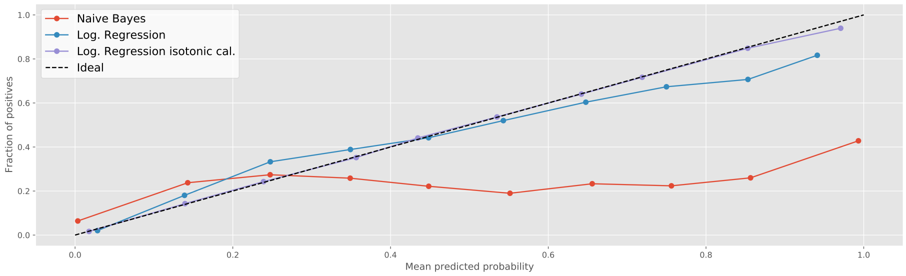

Figure 13: Probability calibration curves.

In the end, I chose not to drop the age feature and stick with 11 out of 16 provided features. I trained logistic regression classifier on all and top 11 features. Just slightly better F-1 score was achieved by using fewer features. Final scores on test data (using 11 of 16 features):

- Accuracy: 0.897,
- Precision: 0.315,
- Recall: 0.625,
- F-1: 0.419.

While F-1 score and precision are slightly lower than our baseline Bayes model's results, the recall was substantially increased from 0.402 to 0.625. We managed to correctly identify $> 60\%$ of all potential customers from the test data sample.

Finally, I tried calibrating probabilities. As mentioned in the Brier score description, we might want to prioritize potential customers, from most likely to subscribe to least. For this reason, we need that our estimated probabilities to be as close as possible to the ”true” probabilities.

Final score:

- Accuracy: 0.896,
- Precision: 0.272,
- Recall: 0.639,
- F-1: 0.381.

Again, whatever we want to calibrate our model depends on business needs. Decreased precision might not be worth a slightly increased recall score (F-1 score decreased compared to uncalibrated model).