Data Mining Assignment 1: Classification

Matteo Vedovati

Understanding the problem

We need to provide a list of potential customers to target for a special promotion. Customers are divided into two different groups:

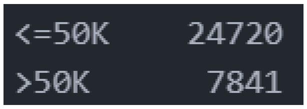
- High-income (more than 50k): 10% likely to accept the offer and will generate 980 euros on average.
- Low-income (less than 50k): 5% likely to accept the offer and will cost 310 euros on average.

The cost of sending the mail to each customer is 10 euros.
The goal is maximizing the revenues meaning that we should only **send the special offer only to the high-income potential customer** because they are the ones that will generate money.
Given that, it’s trivial to understand that this is a binary classification problem (classify potential customers to be high-income or low-income).

The dataset

By analyzing the dataset we can see:

- There are some missing values in some of the features, namely: `workclass`, `occupation` and `native-country`.
- There is an unbalance in the data:



```
<=50K    24720
>50K      7841
```

Preprocessing

- Joined the capital gain and loss features in one single feature.
- For categorical features, I’ve used OneHotEncoder to create a sparse matrix of one-hot encoded columns.
- For handling missing values, given that all the features having missing values were categorical, I've also added a column in the OneHotEncoder for missing values.

Classification methods and evaluations

Given the problem and the dataset, metrics mostly sensitive to false negatives are preferred (the return of finding an actual high-income possible customer is greater than the loss of misclassifying a low-income as a high-income possible customer).
Given this, useful metrics are:

- Recall
- Precision
- F1 score

- AUC-ROC

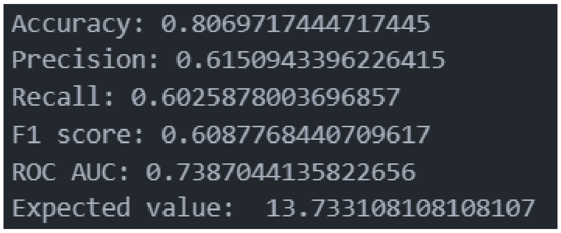
Given the specific problem I created a custom metric that considers the expected value of a potential customer:

$$exp\_val = \frac{n_{over50k}}{n}(precision * 0.1 * 980 - (1 - precision) * 0.05 * 310 - 10)$$

This tailored metric was used to decide what classification method to utilize.

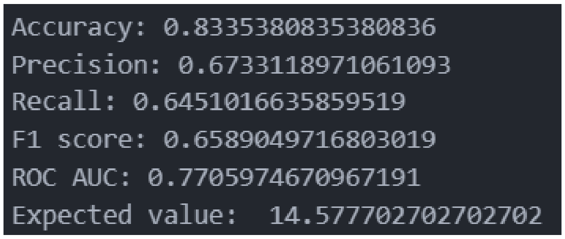
The classification methods tried out and their metrics on the validation set:

Decision Tree:



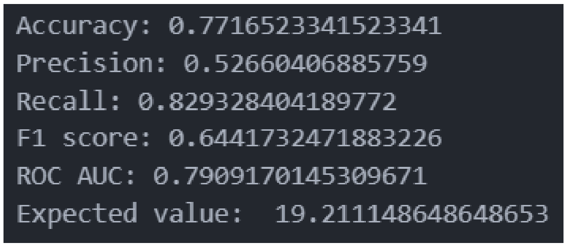
```
Accuracy: 0.8069717444717445
Precision: 0.6150943396226415
Recall: 0.6025878003696857
F1 score: 0.6087768440709617
ROC AUC: 0.7387044135822656
Expected value:  13.733108108108107
```

K neighbors classifier:



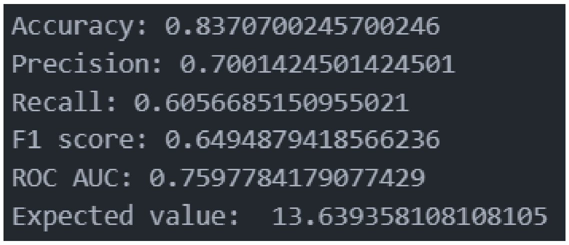
```
Accuracy: 0.8335380835380836
Precision: 0.6733118971061093
Recall: 0.6451016635859519
F1 score: 0.6589049716803019
ROC AUC: 0.7705974670967191
Expected value:  14.577702702702702
```

Naive Bayes:



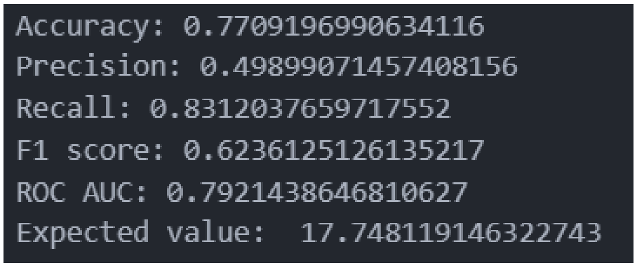
```
Accuracy: 0.7716523341523341
Precision: 0.52660406885759
Recall: 0.829328404189772
F1 score: 0.6441732471883226
ROC AUC: 0.7909170145309671
Expected value:  19.211148648648653
```

Random Forest:



```
Accuracy: 0.8370700245700246
Precision: 0.7001424501424501
Recall: 0.6056685150955021
F1 score: 0.6494879418566236
ROC AUC: 0.7597784179077429
Expected value:  13.639358108108105
```

Given the results on the validation set the classification chosen was the **Naive Bayes classifier**. Metrics results on the test set for the Naive Bayes classifier:

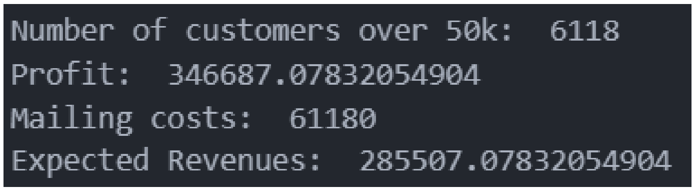


```
Accuracy: 0.7709196990634116
Precision: 0.49899071457408156
Recall: 0.8312037659717552
F1 score: 0.6236125126135217
ROC AUC: 0.7921438646810627
Expected value:  17.748119146322743
```

Results

Give the potential customers:

- The number of classified high-income customers: 6118
- The estimation of the expected gain is: **285,507.08 euros** (while calculating the expected gain we used the precision of the model from the test set for the expected percentage of possible customers classified as high-income to be actually high-income).
  $gain = n_{over50k}(precision * 0.1 * 980 - (1 - precision) * 0.05 * 310 - 10)$



```
Number of customers over 50k:  6118
Profit:  346687.07832054904
Mailing costs:  61180
Expected Revenues:  285507.07832054904
```