Data Glacier

Your Deep Learning Partner

Week 13:
Final presentation for bank marketing campaign project

Name: Rayhanul Islam Rumel
Batch Code: LISUM 17
Submission Date: 30/03/2023
Submission To: Data Glacier

Name: Amogh Vig
Batch Code: LISUM 17
Submission Date: 30/03/2023
Submission To: Data Glacier

Agenda

- Problem Statement
- Approach
- Results
- Final Recommendations

Problem Statement

The Client

ABC Bank wants to sell its term deposit product to customers. Before launching the product, they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

Our Mission

Build a machine learning model that helps ABC Bank shortlist customers whose chances of buying the product is more so that their marketing channel (tele marketing, SMS/email marketing etc) can focus only on those customers.

Approach

1. Used both the bank-full and bank-additional datasets to explore the maximum number of features for our models

2. Performed EDA to identify and fix any problems with the data before feeding this data to the model:
    - Encoding categorical features was a major component of the data processing

3. Implemented different families of machine learning models and fit them to our final data:
    - Ensemble model (e.g., Gradient Boosting)
    - Linear model (e.g., Logistic Regression)
    - Tree-based model (e.g., Random Forest)

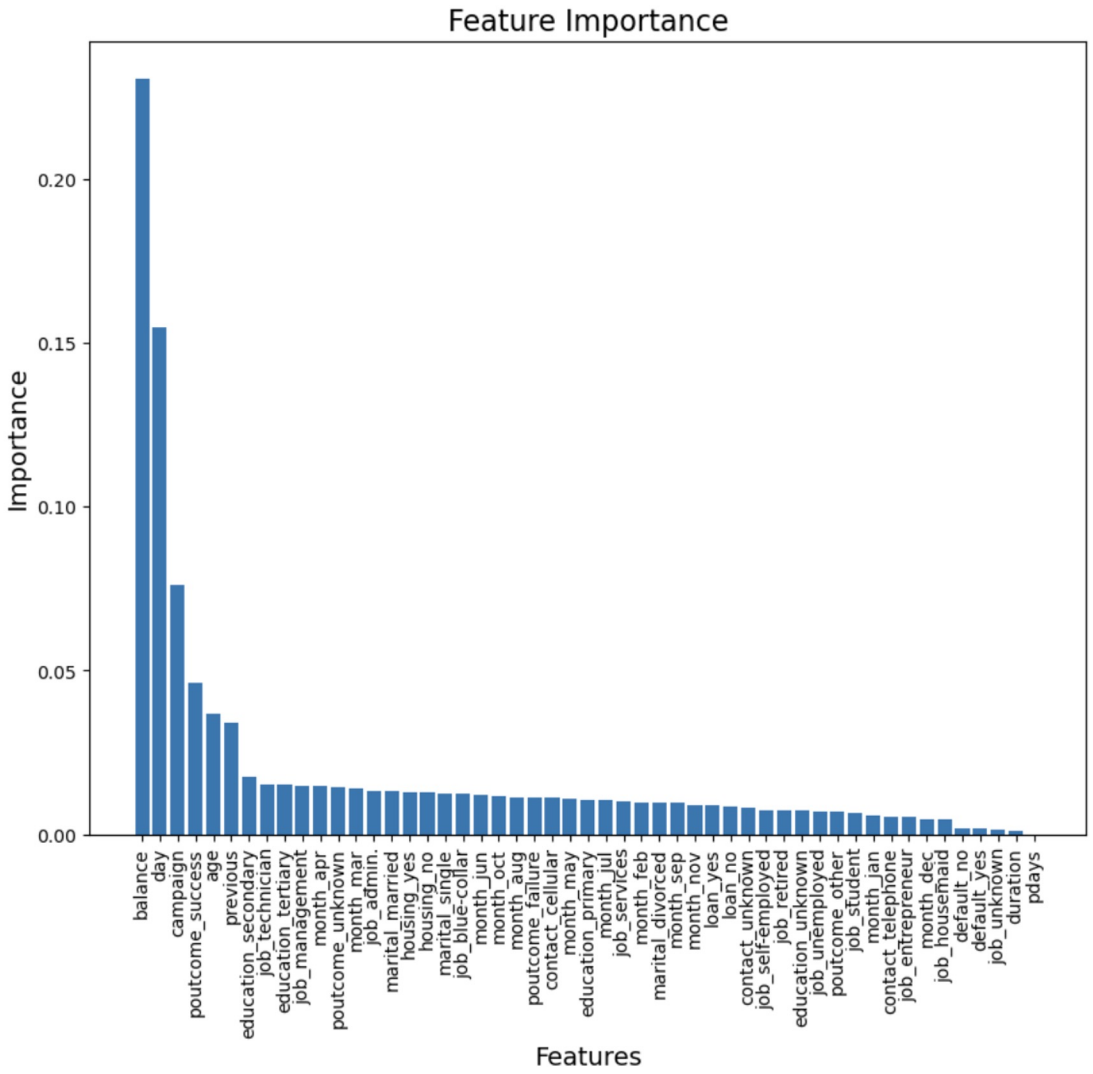
4. Feature selection algorithms were used to identify the most important feature for each dataset

5. Models were evaluated and compared using metrics such as accuracy, precision, recall, and F1-score

Results: Bank-Full Dataset, Feature importance

The most important feature for predicting whether an individual customer will buy the product is their balance, as shown here.

The difference in importance is quite substantial. Most of the other features are quite unimportant in comparison to the balance variable.



Results: Bank-Full Dataset, ML Models

Many different models were instantiated

For preliminary model testing, the mean cross-validation scores were computed using the training set to identify top candidates. The results are shown in the table below:

	Classifier	Crossval Mean Scores
4	Grad Boost CLF	0.893768
5	Random Forest	0.886772
0	Logistic Reg.	0.884118
1	SVC	0.883012
2	KNN	0.874440
3	Dec Tree	0.829702
6	Neural Classifier	0.824723
7	Naives Bayes	0.824723

Based on these results, the top 3 most promising candidates were Gradient Boosting Classifier, Random Forest Classifier, and Logistic Regression
These 3 models were therefore selected to be fit to the data and evaluated using the test set

Results: Bank-Full Dataset, ML Models

Gradient Boosting Classifier

	precision	recall	f1-score	support
no	0.90	0.99	0.94	7985
yes	0.69	0.21	0.32	1057
accuracy			0.90	9042
macro avg	0.80	0.60	0.63	9042
weighted avg	0.88	0.90	0.87	9042

Logistic Regression

	precision	recall	f1-score	support
no	0.88	1.00	0.94	7985
yes	0.50	0.00	0.01	1057
accuracy			0.88	9042
macro avg	0.69	0.50	0.47	9042
weighted avg	0.84	0.88	0.83	9042

Random Forest Classifier

	precision	recall	f1-score	support
no	0.91	0.98	0.94	7985
yes	0.61	0.24	0.35	1057
accuracy			0.89	9042
macro avg	0.76	0.61	0.64	9042
weighted avg	0.87	0.89	0.87	9042

Results: Bank-Full Dataset, ML Models

After considering all evaluation metrics, Gradient Boosting Classifier is the best candidate

Reasons:

- Highest weighted average F1-score of 0.87 (tied with Random Forest)
- Highest accuracy of 90%

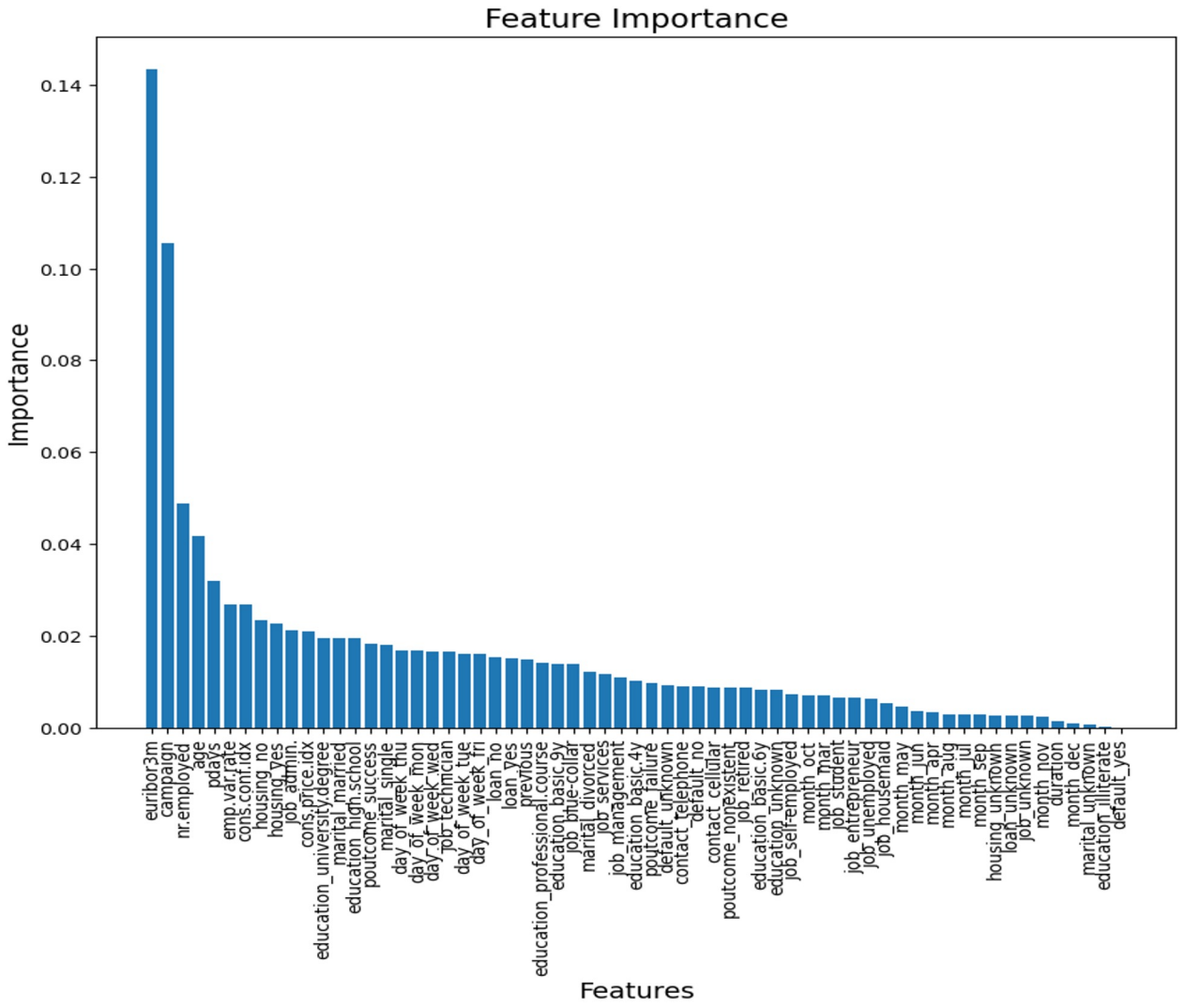
Note that our main evaluation metric is weighted average F1-score:

- Weighted average F1-score is a better metric than accuracy for our use case, due to class imbalances in the data
- Thus, since it has both the tied highest weighted average F1-score and the highest accuracy, we can conclude that Gradient Boosting Classifier is the best model

Results: Bank-Addition-Full Dataset, Feature importance

The most important feature for predicting whether an individual customer will buy the product is their euribor3m, as shown here. The next most significant feature is campaign.

Most of the other features are quite unimportant in comparison to the euribor3m, and campaign.



Results: Bank-Additional-Full Dataset, ML Models

Two different ML models were instantiated: Random Forest Classifier, and Logistic Regression

Random Forest Classifier

```
Accuracy: 0.8838169236372466
Precision: 0.49070631970260226
Recall: 0.2787750791974657
F1 Score: 0.8694024263695561
```

Logistic Regression

	precision	recall	f1-score	support
no	0.91	0.98	0.94	7290
yes	0.59	0.22	0.32	947
accuracy			0.89	8237
macro avg	0.75	0.60	0.63	8237
weighted avg	0.87	0.89	0.87	8237

```
f1_score(y_test, y_pred, pos_label='yes', average='weighted')
```

```
0.8708025489748393
```

Results: Bank-Additional-Full Dataset, ML Models

After considering all evaluation metrics, Logistic Regression is the best candidate

Reasons:

- Highest F1-score of 87% where for Random Forest Classifier the FI-score is 86.9%
- The Highest accuracy of 89% (even though only slightly higher than Random Forest Classifier, which has an accuracy of 88.38%)

Note that our main evaluation metric is F1-score:

- F1-score is a better metric than accuracy for our use case, due to class imbalances in the data
- Thus, since Logistic Regression has both the highest F1-score and accuracy, we can conclude that Logistic Regression is the best model

Recommendations

Final comments

- For Bank-Full dataset, we have shown that Gradient Boosting Classifier is the best candidate among all the models tested, with a weighted average F1-score of 0.87
- For Bank-Additional-Full dataset, we have shown that Logistic Regression is the best candidate among the models considered in the experiment, with the highest recorded F1-score of 0.87 where the F1-score for Random Forest Classifier was 0.869
- In this study, the models were implemented using default hyperparameters

Final recommendation

- ABC Bank should select Gradient Boosting Classifier (for Bank-Full dataset) and Logistic Regression (for Bank-Additional-Full dataset) as their baseline models for predicting whether a given customer will purchase their product
- ABC Bank should perform hyperparameter tuning using methods such as GridSearchCV and/or RandomizedSearchCV to improve the model from the baseline by achieving a higher weighted average F1-score

Thank You