MACHINE LEARNING OPERATIONS

REPORT

2022/2023

Afonso Reyna 20191197
Evans Onorieru
Mariana Rodrigues 20220627
Samuel Santos 20220609

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Github link:

1. Motivation

The motivation for choosing the Bank Marketing dataset for this project lies in its relevance to real-world scenarios and the opportunity to simulate the process of deploying machine learning models. This dataset also offers a practical and challenging problem domain.

Contextualization:

The dataset represents direct marketing campaigns conducted by a Portuguese banking institution. By working with this dataset, we can gain insights into the factors that influence customers' decision to subscribe to a term deposit. This closely resembles a common business problem faced by banks and financial institutions, making it a suitable choice for simulating real-world scenarios.

2. Success Metrics

Subscription Conversion Rate: We used this metric to get the percentage of individuals who made a subscription out of the total number of individuals targeted. It indicates the effectiveness of your marketing or sales efforts in converting leads into subscribers.

- We obtained a conversion rate of approximately 0.97, indicating that a high proportion of the instances in the dataset have a positive prediction (1) in the 'prediction' column. Out of the total instances, there are 21320 instances with a prediction value of 1 and 663 instances with a prediction value of 0.

3. Project Planning

- Sprint 1: 2 days: The goal was to choose the dataset, explore the data, taking some interesting conclusions and build the unit tests pipeline.
- Sprint 2: 2 days: The goal was to create the clean data and feature engineering steps, build the data preprocessing pipeline, and build the data split pipeline.
- Sprint 3: 2 days: The goal was to build the model train, model predict, and feature selection pipelines.
- Sprint 4: 2 days. The goal was to explore data drift and create its pipeline.
- Sprint 5: 1 day: Analyse the results and report consolidation.

4. Results and Conclusions:

4.1. Data Exploration

This exploratory analysis aims to provide insights into our dataset which contains 17 variables capturing various aspects of the individuals and their interactions with the marketing campaigns. These variables include demographic information, such as age, marital status, education, credit status, financial balance, and loan information. We also have some information about the contact method,

timing, and outcome of previous marketing campaigns as well as whether the individuals subscribed to the product or service.

Our analysis involves exploring the relationships and patterns within the dataset, allowing us to gain valuable information about the target audience and their performances. We will conduct various data exploration techniques, including descriptive statistics and visualizations.

- Based on the **correlation matrix**, we took some conclusions:
    - Age and Subscription have a weak correlation (0.025), suggesting that older customers may have a slightly higher likelihood of subscribing to services.
    - Balance (euros) and Subscription have a weak correlation (0.053), suggesting that customers with higher account balances are more likely to subscribe.
    - Last Contact Day and Subscription show a negative correlation (-0.028), indicating that the specific day of the month on which a customer was last contacted has a minimal influence on their likelihood of subscribing.
    - Campaign and Subscription have also a weak negative correlation (-0.073), implying that excessive or repeated contacts within a campaign may hurt the subscription rate.
    - Pdays and Subscriptions have a weak positive correlation (0.104), suggesting that a longer duration since the last contact increases the chances of a customer subscribing in the current campaign.
    - Previous and Subscription have also a positive correlation (0.093), indicating that customers who were contacted more frequently in previous campaigns are more likely to subscribe in the current campaign.
- Regarding ***'Subscription'*** distribution, it was possible to conclude that the majority represented the subscriptions made, being 88.30% of the dataset, while no subscription (2) accounts for a smaller proportion 11.70%, suggesting an imbalance in the distribution of subscription outcomes.
- There are no **missing values,** but we have a lot of strange records with 'unknown'.
    - 288 unknown *Jobs*;
    - 13020 unknown *Contacts*;
    - 36959 unknown *Poutcomes*;
    - 1857 unknown *Education* levels;

  There is a huge presence of 'unknown' values in certain features in the dataset.
- The ***'balance'*** is generally higher for subscriptions that were made compared to subscriptions that were not made, concluding that there appears to be a positive correlation between the balance (euros) and the likelihood of making a subscription because as the balance increases, the probability of making a subscription tends to decrease. This suggests that individuals with higher financial capacity may be less inclined to subscribe.
- Regarding information about the ***'Job'*** types of individuals, the most common job categories are "management" (9732 individuals), followed by "technician" (9458 individuals) and "entrepreneur" (7597 individuals).
- Among the records where the ***'Job'*** has not an 'unknown' Subscription, there are 39668 subscriptions made and 5255 not made. Regarding the records where the 'Job' is labelled as 'unknown', there are 254 Subscriptions and 34 where the subscription was not made. It indicates that most of the rows with the 'unknown' job values subscribed.
- ***'Marital Status'*** distribution indicates that the most frequent marital status is married, followed by single and divorced, with 27214, 12790, and 5207 records, respectively. Among

married individuals, there were 24459 subscriptions made, while 2755 instances represent no subscription made, indicating that a significant number of married individuals subscribed to the offered services. Single individuals also showed a considerable level of subscription adoption, with 10878 subscriptions and 1912 not subscribed, suggesting that both subscription options were chosen by a significant number of single individuals. Even within the divorced category, there was a notable level of subscription adoption.

- Regarding ***'Education'***, secondary is the most frequent education level, with 23202 individuals, while tertiary, primary, and unknown have 13301, 6851, and 1857, respectively. There were always more subscriptions than non-subscription across all education levels.
- The conclusions drawn about the ***'Poutcome'*** variable were that the majority of the instances have an 'unknown' outcome of the previous marketing campaign, with a count of 36959. There are 4901 instances where the previous campaign was a 'failure', while 1511 was a 'success' and the remaining were categorized as 'other'. The presence of many 'unknowns' suggests that there might be missing data for the outcome of previous campaigns. Additionally, note that varying frequencies of different outcome categories, such as 'failure', 'other', and 'success' could indicate the effectiveness of the previous marketing efforts.
- Regarding individuals with and without ***'housing loans'***, we can conclude that the dataset contains more individuals with a housing loan, accounting for 25130, which represents 55.6%, while individuals without housing loans represent 44% of the dataset.
- The majority of the individuals do not have a ***'credit'***, representing 44396 individuals. It is also possible to see that most subscriptions were made by individuals without credit defaults, suggesting that individuals who do not have credit defaults are more likely to subscribe. Also, there are relatively low subscriptions made by individuals with credit default, with 763 individuals subscribing, indicating that individuals with credit are less likely to subscribe compared to those without credit.

4.2. Data Modelling

Feature Importance:

The methodology used to determine the importance of features was Recursive Feature Elimination (RFE). RFE assigns a ranking to each feature, indicating its relative importance. Features with higher rankings are considered more influential in determining the target variable. Based on results, we have identified the top-ranked features, that are represented in yellow in the following table.

Selected	
Feature	**Ranking**
age	1
education_secondary	1
marital_status_married	1
job_technician	1
previous	1
pdays	1
campaign	1
job_management	1
last_contact_month	1
balance_euros	1
housing_loan	1
last_contact_duration	1
contact_unknown	1
last_contact_day	1

Discarded	
Feature	**Ranking**
education_tertiary	2
personal_loan	3
job_blue-collar	4
marital_status_single	5
contact_telephone	6
job_services	7
education_unknown	8
job_retired	9
job_student	10
job_unemployed	11
job_self-employed	12
job_entrepreneur	13
job_housemaid	14
credit	15
job_unknown	16

Therefore, the inclusion of only the selected features in our analysis ensures a more focused approach, enabling us to derive meaningful insights and make accurate predictions based on the most influential variables.
We reassessed the feature importance in the champion model, and we got these results:

Feature	Importance
last_contact_duration	0.447
last_contact_month	0.092
pdays	0.089
age	0.077
housing_loan	0.062
last_contact_day	0.057
balance_euros	0.052
previous	0.044
contact_unknown	0.026
campaign	0.021
personal_loan	0.009
job_blue-collar	0.009
education_tertiary	0.008
job_student	0.007

Explainability:

Regarding explainability of our model, we decided to use models, such as Random Forest that inherently offer high explainability because they provide clear coefficients that can explain the importance of each feature in prediction. We also used techniques like SHAP for feature importance analysis to understand the contribution of each feature to the model´s predictions, allowing us to explain and implement the model´s behaviour.

5. Discuss POC

Our setup is based on python, supported by kedro, mlflow, great expectations and nannyml, as long as the usual python libraries, like Pandas for data processing and analysis because of its ease of use, rich functionality and because of our small data volume. Pandas also allowed us to quickly prototype and validate our proof of concept, enabling efficient data manipulation and exploration.
Our project establishes a series of data-handling processes that are then declaratively organized in pipelines. Each of these processes is essentially independent, with the inputs and outputs mapped to values defined in a catalog. We defined a CHAMPION model and several CHALLENGER models, and we can easily compare the results, using shadow testing. Each model must be fitted with the train data and then evaluated on the daily data (the test data). Unit tests are used to ensure that the data is compliant with the requirements; for this we use great expectations. For data drift, we use a nanny ml univariate analysis (for simplicity, we chose just a tiny number of features).
A risk to consider is the availability of computational resources. Depending on the infrastructure and hardware available in a production environment, the computational requirements of our models and data preprocessing pipeline may exceed the available resources. To mitigate this risk, the process may involve optimizing memory usage, parallelizing computations, or leveraging solutions to scale resources needed.

6. Packages and versions

databricks-cli==0.17.7
docker==6.1.2
evidently==0.3.3
exceptiongroup==1.1.1
gitdb==4.0.10
GitPython==3.1.31
graphviz==0.20.1
great-expectations==0.16.16
greenlet==2.0.2
ipython @ file:///D:/bld/ipython_1685727936079/work
ipython-genutils==0.2.0
ipywidgets==8.0.6
kedro==0.18.8
kedro-datasets==1.4.1
kedro-mlflow==0.11.8
Markdown==3.4.3
matplotlib==3.7.1
matplotlib-inline @ file:///home/conda/feedstock_root/build_artifacts/matplotlib-
mlflow==2.4.1
nannyml==0.8.6
numpy==1.23.1
pandas==1.5.3
pandas-profiling==3.3.0
plotly==5.15.0
pytest==7.3.1
pytest-cov==4.0.0
scikit-learn==1.2.2
scipy==1.10.1
seaborn==0.11.2
shap==0.41.0
slicer==0.0.7
statsmodels==0.14.0
tqdm==4.65.0
xgboost==1.7.6