Week 11 - Explain the model, analyze risk, bias and ethical considerations

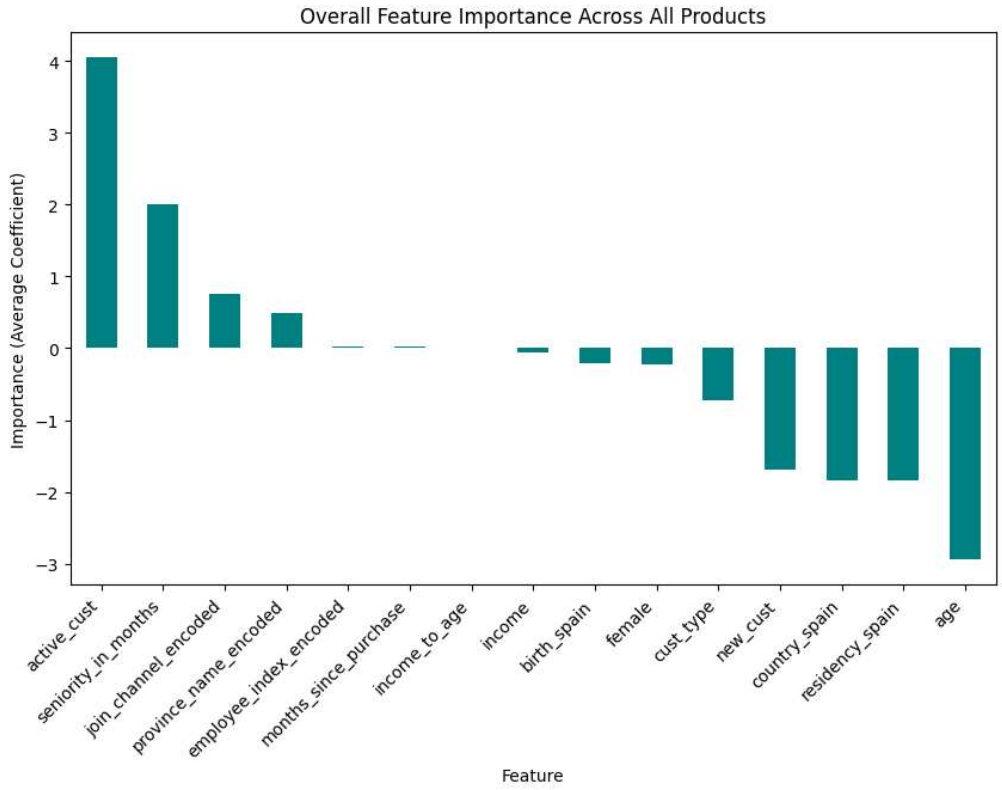
Aritra Ray

Maria Alice Fagundes Vieira

- **Identification of important features in your model. What are the top 5-10 predictors?**

Since our model is a logistic regression, it is a glass box model as it provides clear, interpretable coefficients for each feature, indicating the relationship between the feature and the target variable and we can directly examine feature importance using the coefficients (or their absolute values).

We used a coef_[0] code to generate the importance coefficient values for the variables. The result returned as below:



Top Positive Contributors:

- active_cust shows the highest positive importance (approximately +4.0)
- seniority_in_months is the second most important feature (around +2.0)

- join_channel_encoded and province_name_encoded show moderate positive influence (between 0.5-1.0)

Neutral Features

Several features show minimal impact near zero:

- employee_index_encoded
- months_since_purchase
- income_to_age
- income

Top Negative Contributors:

- age shows the strongest negative correlation (approximately -3.0)
- residency_spain and country_spain have significant negative importance (around -2.0)
- new_cust and cust_type also show notable negative influence

The distribution of feature importance shows a hierarchy, with customer activity metrics being most influential, followed by tenure-related features, while demographic and location-based features tend to have negative correlations with the target variable. This suggests the model relies heavily on behavioral and relationship metrics rather than static customer attributes for making predictions. While this is great, as it decreases bias, the demographic information is still important in the prediction model, and I for some variables more than others, (as we will see on the LIME analysis) which introduces bias. However, overall, the model performs well and is generically unbiased.

We proceeded to do a global vs local model interpretation. Global interpretation explains the overall behavior of the model and is ideal for understanding how features generally affect predictions across the dataset. Since logistic regression coefficients inherently provide global interpretability, we ran a local Interpretation for one of the products (savings_acct), to understand individual predictions, and it is useful to understand why the model made a specific decision for a single instance.

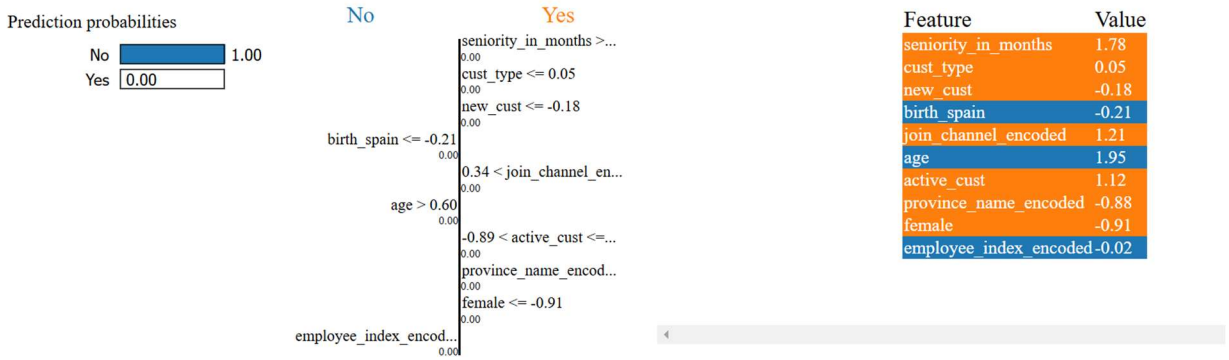
To run a local interpretation, we decided to use LIME because it is quick, simpler and faster. LIME is model-agnostic, meaning it can be used to explain predictions of any machine learning model,

regardless of the algorithm used. It works well when interpretability for specific predictions is needed, which is what we will do with the savings_acct variable.

The LIME output highlights the contribution of each feature for the specific prediction ('savings_acct'):

- Positive contributions (towards "Yes" in this case) are in green.
- Negative contributions (towards "No") are in red.
- It shows which features were most influential in the prediction.

We ran the LIME code and got the below results for savings_acct:



The right side shows a table and visualization of feature importance and their values

Most Influential Positive Features:

- Age (1.95) - strongly pushes toward the prediction
- Seniority in months (1.78) - second most important positive factor
- Join channel encoded (1.21) - contributes positively
- Active customer (1.12) - also has a positive influence

Most Influential Negative Features:

- Female (-0.91) - pushes against the prediction
- Province name encoded (-0.88) - negative contribution
- Birth Spain (-0.21) - slight negative influence

- New customer (-0.18) - minimal negative impact

Decision Rules

The middle section shows the decision rules that led to this prediction, with conditions like:

- Seniority in months > certain threshold
- Customer type <= 0.05
- Birth Spain <= -0.21
- Age > 0.60

The values shown represent the feature contributions to the model's decision, where positive values push toward one class and negative values push toward the other. The magnitude of these values indicates the strength of their influence on the prediction.

- **Select 5 predictions at random, explain how the model generated those predictions (which features matter more than others), which features need to change and by how much to move the output in a significant way (e.g., to flip the prediction from one class to another)**

  After running the 5 predictions at random, the model only suggested me to change the 'active_cust' from 0 to 1 or the opposite since it is the most influential variable, which makes sense, since if a client is not an active client they will not buy any product and if the client is an active client there is a chance of the client buying the product. Then other variables enter into play such as customer type, join channel, employee index and new customer.

- **Discuss whether your dataset includes protected categories and whether you are using them in your model. Discuss the bias of the model (with respect to the protected categories if your dataset includes them, otherwise talk about the correlations between your variables and protected categories)**

  Our dataset includes protected categories such as gender, age and national origin and we are using those because they can be an important variable to predict the product a client will buy. Based on the LIME results for savings_acct, these protected variables have significant influence on the model's decisions - age shows the highest positive impact (1.95), while being female has a substantial negative effect (-0.91), and birth location (Spain) contributes negatively (-0.21).

For instance, age could be a factor for a young client, which will be unlikely to buy a wealth management product, since they probably don't have enough income/savings, and will be more likely to buy the junior account or a savings account product, since they are building their wealth or starting to save their income. This affects our model since the bank may choose not to direct their resources to promote a product that is unlikely for a young client to buy, and instead, they can offer something that young clients are more likely to buy.

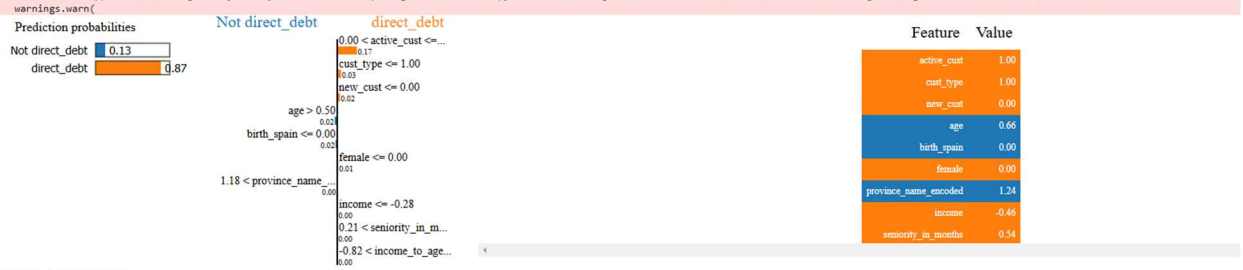
- **Provide bias removal strategies and their impact on the predictions**

To address these potential biases, several strategies can be implemented. First, protected attributes could be removed entirely from the model or modified - for example, using age ranges instead of exact age to reduce age discrimination. Second, the model could be modified through fairness constraints during training or by implementing adversarial debiasing techniques. Third, alternative features focusing on behavior-based metrics, transaction history, and product usage patterns could replace demographic variables. While these changes might slightly reduce model accuracy, they would promote more equitable treatment across demographic groups, reduce regulatory risks, and create more sustainable and ethical business practices.

APPENDIX

Random prediction 1:

```
Explaining Prediction 1 (Row 950916):
Predicted Class: 1 (Probability: 0.87)
Intercept -0.039240922130545156
Prediction_local [0.15935189]
Right: 0.8716594547837396
c:\Users\MARIA\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with feature names
  warnings.warn(
c:\Users\MARIA\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with feature names
  warnings.warn(
```

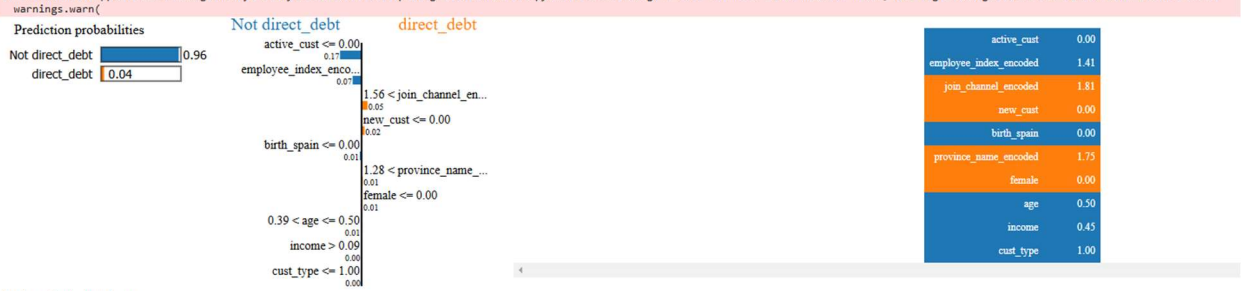



```
Feature Contributions:
0.00 < active_cust <= 1.00: 0.1687
cust_type <= 1.00: 0.0320
new_cust <= 0.00: 0.0229
age > 0.50: -0.0219
birth_spain <= 0.00: -0.0151
female <= 0.00: 0.0084
1.18 < province_name_encoded <= 1.28: -0.0037
income <= -0.28: 0.0034
0.21 < seniority_in_months <= 0.54: 0.0023
-0.82 < income_to_age <= -0.35: 0.0016

To flip the prediction, try adjusting '0.00 < active_cust <= 1.00' by a significant amount.
```

Random prediction 2:

```
Explaining Prediction 2 (Row 3239):
Predicted Class: 0 (Probability: 0.96)
Intercept 0.21402389014112455
Prediction_local [0.03316361]
Right: 0.03768068101520936
c:\Users\MARIA\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with feature names
  warnings.warn(
c:\Users\MARIA\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with feature names
  warnings.warn(
```

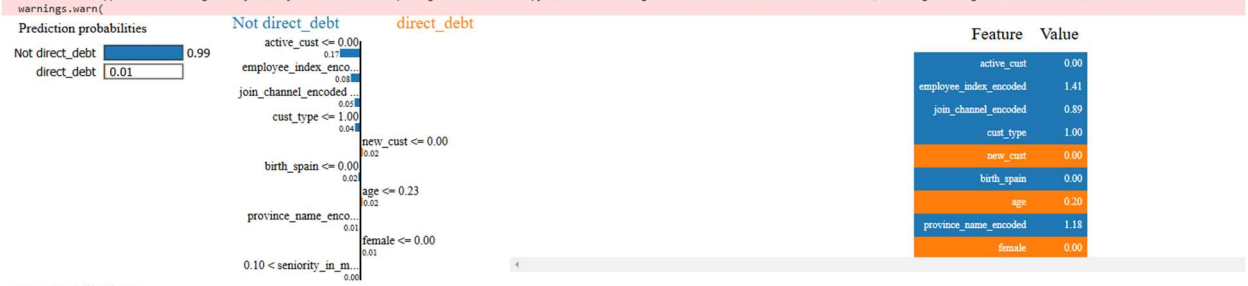



```
Feature Contributions:
active_cust <= 0.00: -0.1708
employee_index_encoded <= 1.41: -0.0710
1.56 < join_channel_encoded <= 1.94: 0.0472
new_cust <= 0.00: 0.0232
birth_spain <= 0.00: -0.0136
1.28 < province_name_encoded <= 1.75: 0.0093
female <= 0.00: 0.0069
0.39 < age <= 0.50: -0.0056
income > 0.09: -0.0043
cust_type <= 1.00: -0.0021

To flip the prediction, try adjusting 'active_cust <= 0.00' by a significant amount.
```

Random prediction 3:

```
Explaining Prediction 3 (Row 458596):
Predicted Class: 0 (Probability: 0.99)
Intercept 0.3001528352611889 7
Prediction_local [-0.02265368]
Right: 0.014329623063726801
c:\Users\MARIA\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with feature names
  warnings.warn(
c:\Users\MARIA\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with feature names
  warnings.warn(
```



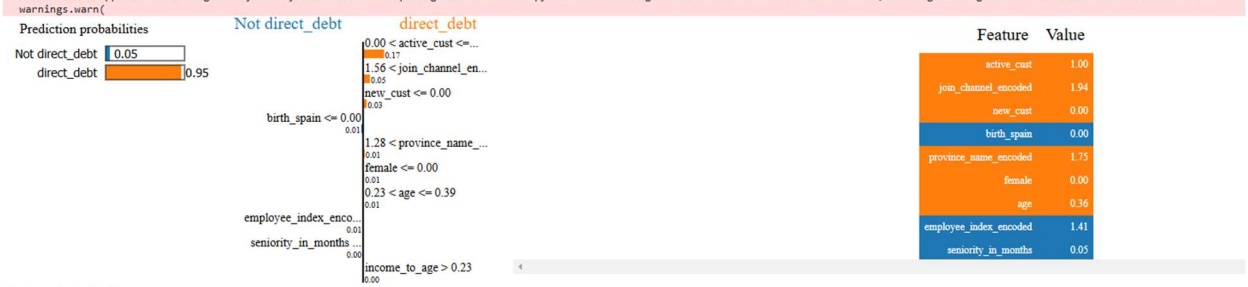

```
Feature Contributions:
active_cust <= 0.00: -0.1668
employee_index_encoded <= 1.41: -0.0780
join_channel_encoded <= 0.89: -0.0475
cust_type <= 1.00: -0.0443
new_cust <= 0.00: 0.0195
birth_spain <= 0.00: -0.0178
age <= 0.23: 0.0173
province_name_encoded <= 1.18: -0.0074
female <= 0.00: 0.0067
0.10 < seniority_in_months <= 0.21: -0.0044

To flip the prediction, try adjusting 'active_cust <= 0.00' by a significant amount.
```

Random prediction 4:

```
To flip the prediction, try adjusting 'active_cust <= 0.00' by a significant amount.

Explaining Prediction 4 (Row 1115023):
c:\Users\MARIA\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with feature names
  warnings.warn(
Predicted Class: 1 (Probability: 0.95)
Intercept -0.03681746983404906
Prediction_local [0.22245744]
Right: 0.9472943668943028
c:\Users\MARIA\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with feature names
  warnings.warn(
```

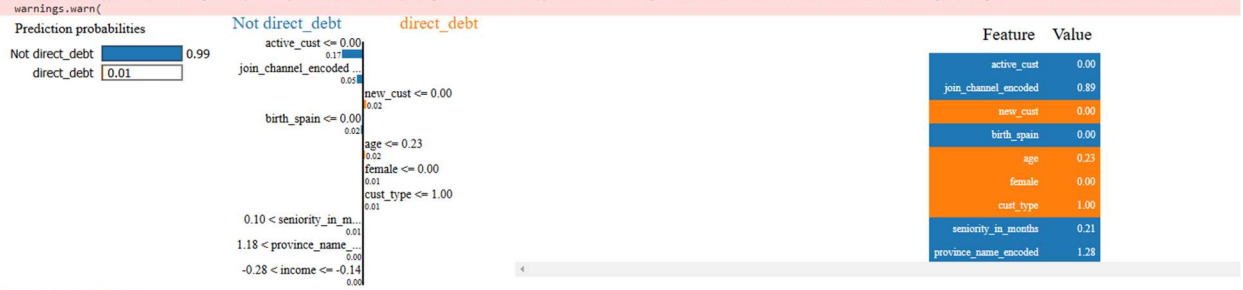



```
Feature Contributions:
0.00 < active_cust <= 1.00: 0.1695
1.56 < join_channel_encoded <= 1.94: 0.0546
new_cust <= 0.00: 0.0291
birth_spain <= 0.00: -0.0150
1.28 < province_name_encoded <= 1.75: 0.0127
female <= 0.00: 0.0079
0.23 < age <= 0.39: 0.0071
employee_index_encoded <= 1.41: -0.0055
seniority_in_months <= 0.10: -0.0045
income_to_age > 0.23: 0.0035

To flip the prediction, try adjusting '0.00 < active_cust <= 1.00' by a significant amount.
```

Random prediction 5:

```
Explaining Prediction 5 (Row 557414):
Predicted Class: 0 (Probability: 0.99)
c:\Users\MARIA\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with feature names
  warnings.warn(
Intercept 0.1629545436552834
Prediction_local [-0.02360619]
Right: 0.014531659456663631
c:\Users\MARIA\AppData\Local\Programs\Python\Python311\Lib\site-packages\sklearn\base.py:464: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with feature names
  warnings.warn(
```




```
Feature Contributions:
active_cust <= 0.00: -0.1665
join_channel_encoded <= 0.89: -0.0483
new_cust <= 0.00: 0.0249
birth_spain <= 0.00: -0.0164
age <= 0.23: 0.0161
female <= 0.00: 0.0079
cust_type <= 1.00: 0.0074
0.10 < seniority_in_months <= 0.21: -0.0063
1.18 < province_name_encoded <= 1.28: -0.0042
-0.28 < income <= -0.14: -0.0012

To flip the prediction, try adjusting 'active_cust <= 0.00' by a significant amount.
```