Assignment 5.1 Exercise

SHAP Loan Approval Explainability Analysis

In this exercise, you will analyze the loan approval dataset and utilize SHAP to understand feature contributions to model predictions. The goal is to interpret how specific features affect decisions and reflect on the ethical implications of such models.

Instructions:

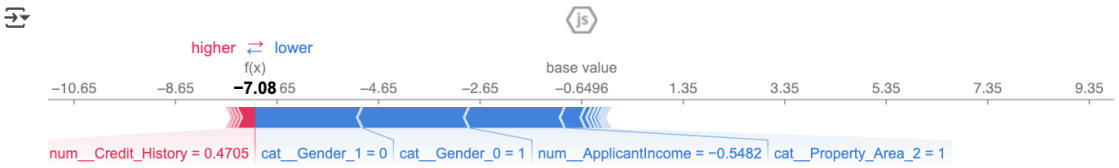
1. Complete “Step 3: SHAP Explanations” in the **assignment_5.1_notebook** copy and paste your output and respond to the questions into this exercise document.
2. Provide short responses to the reflection questions in “Step 4: Ethical Reflection”

Step 3: SHAP Explanations

Use SHAP to explain the predictions of your logistic regression model.

1a. Generate the SHAP force plot for the first data point in your test set using the SHAP values. The force plot should explain how individual features contribute to pushing the model’s prediction towards or away from class 1.

Answer:



Interpretation of the Plot Results

The SHAP force plot explains why the model predicted a very low probability of loan approval for this specific applicant.

The model’s output (f(x) = -7.08) is much lower than the base value (-0.65), indicating a strong prediction for Class 0 (loan rejection). Several features contributed to pulling the prediction downward, including the applicant being male (cat__Gender_0 = 1), having a low income (num__ApplicantIncome = -0.5482), and residing in a certain property area (cat__Property_Area_2 = 1).

The only factor that slightly pushed the prediction upward was the applicant’s credit history (num__Credit_History = 0.4705). Overall, the combined impact of the features resulted in a prediction with less than 0.1% probability of loan approval.

1b. Which features are pushing the prediction towards class 1? Which features are pushing the prediction away from class 1?

Response:

The features pushing the prediction towards class 1 and away from class 1 are:

```
 Features pushing TOWARDS ----> class 1:
num__Credit_History: +0.3552
cat__Dependents_0: +0.0733
num__CoapplicantIncome: +0.0681
cat__Dependents_1: +0.0573
cat__Dependents_3: +0.0204

 Features pushing AWAY -----> class 1:
cat__Gender_1: -2.1069
cat__Gender_0: -2.1058
num__ApplicantIncome: -1.8854
cat__Property_Area_2: -0.4069
cat__Property_Area_1: -0.1174
cat__Property_Area_0: -0.0828
cat__Married_1: -0.0724
cat__Married_0: -0.0715
cat__Dependents_2: -0.0548
num__LoanAmount: -0.0346
num__Loan_Amount_Term: -0.0303
cat__Self_Employed_1: -0.0194
cat__Self_Employed_0: -0.0192
```

```
Indices of class 0:


Index([ 17,  23,  37,  45,  48,  49,  50,  51,  52,  54,
       ...
       603, 604, 605, 606, 608, 609, 610, 611, 612, 613],
      dtype='int64', length=394)

Indices of class 1:


Index([  0,   1,   2,   3,   4,   5,   6,   7,   8,   9,
       ...
       498, 530, 548, 556, 562, 581, 588, 592, 599, 607],
```

```
dtype='int64', length=220)
```

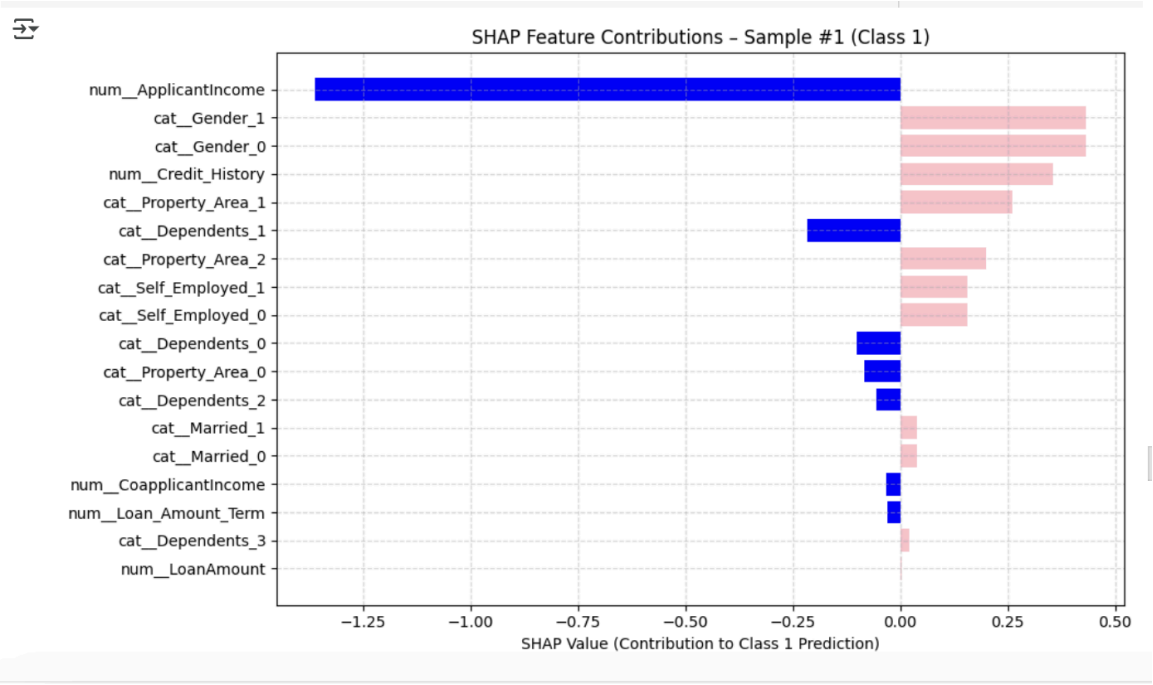
The features pushing the prediction towards class 0 and away from class 0 are:

```
Features pushing TOWARDS ----> class 0 (for sample 0 ):
num__Credit_History (value: 0.4705): +0.3552
cat__Dependents_0 (value: 1.0000): +0.0733
num__CoapplicantIncome (value: -0.5357): +0.0681
cat__Dependents_1 (value: 0.0000): +0.0573
cat__Dependents_3 (value: 0.0000): +0.0204

Features pushing AWAY -----> class 0 (for sample 0 ):
cat__Gender_1 (value: 0.0000): -2.1069
cat__Gender_0 (value: 1.0000): -2.1058
num__ApplicantIncome (value: -0.5482): -1.8854
cat__Property_Area_2 (value: 1.0000): -0.4069
cat__Property_Area_1 (value: 0.0000): -0.1174
cat__Property_Area_0 (value: 0.0000): -0.0828
cat__Married_1 (value: 0.0000): -0.0724
cat__Married_0 (value: 1.0000): -0.0715
cat__Dependents_2 (value: 0.0000): -0.0548
num__LoanAmount (value: -0.8952): -0.0346
num__Loan_Amount_Term (value: 0.3274): -0.0303
cat__Self_Employed_1 (value: 0.0000): -0.0194
cat__Self_Employed_0 (value: 1.0000): -0.0192
```

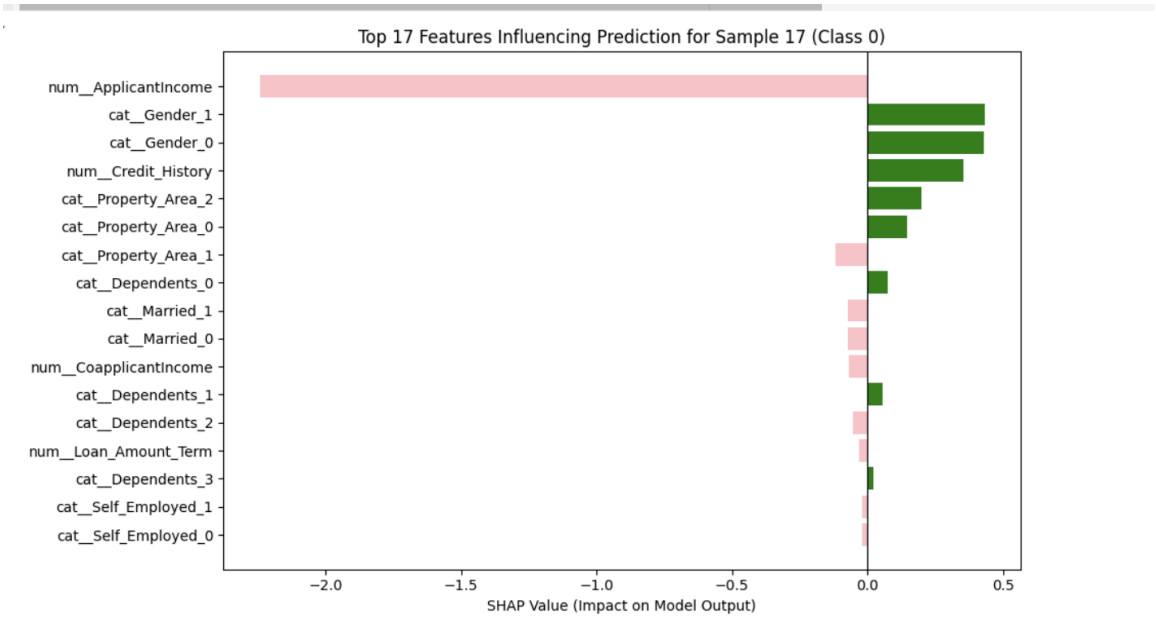
2a. Create a bar plot that displays the SHAP values (feature contributions) for the second data point and SHAP values for class 1. The bar plot will show how much each feature contributes to the model's prediction for class 1, with the features ranked by their contribution.)

Answer:

The plot displays the feature of the second data point and SHAP values for Class 1



The plot displays the feature of the 17th data point and SHAP values for Class 0



2b. Which feature had the greatest influence on the prediction for this data point? Were most features pushing the prediction toward or away from class 1? Are there any surprising variables that are influencing the prediction?

Response:

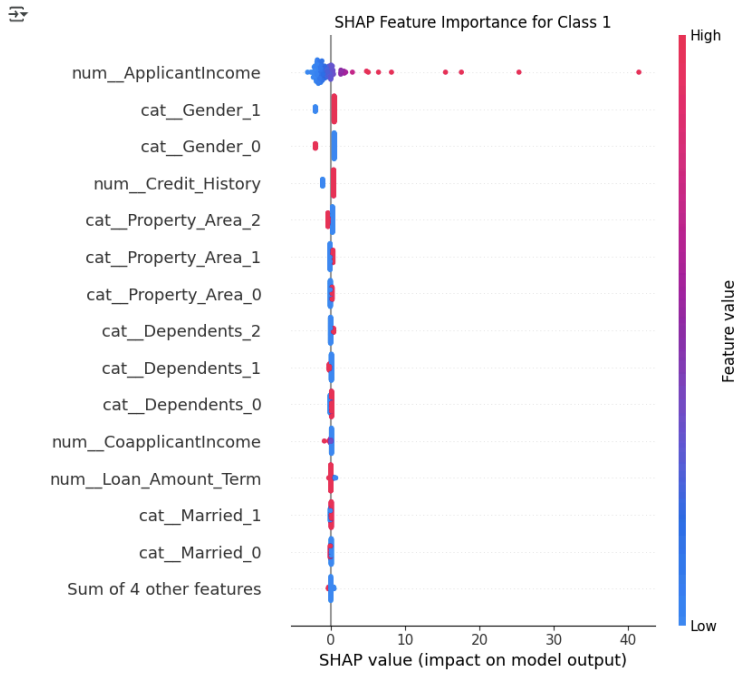
```
The feature with the greatest influence is: num__ApplicantIncome
It is pushing the prediction away from class 1.

Features pushing toward class 1: 11
Features pushing away from class 1: 7

Surprising features (e.g., non-intuitive or weak predictors with high impact):
- num__ApplicantIncome with SHAP value -1.3610
```

Results :

The model's prediction was strongly influenced by the feature `num__ApplicantIncome`, which had the highest impact and pushed the **prediction away from Class 1**, likely contributing to a **negative outcome such as loan rejection.** While 11 features supported Class 1 (approval), only 7 opposed it, yet the negative influence of **ApplicantIncome outweighed the positive contributions.** Due to its unexpectedly large SHAP value (–1.3610), `num__ApplicantIncome` is also flagged as a **surprising feature**, indicating it played a disproportionately strong role in the model's decision.

3a. Create a beeswarm plot to visualize the global importance of features.

Answer Beeswarm plot visualizing the global importance of features:



3b. Which feature appears to be the most important across all data points? Are there any features that have a mixed effect (i.e. pushing predictions towards class 1 and other times pushing them away?) What could cause this variability? How do the color of the dots help you

interpret the relationship between the feature value and the prediction? Are there any notable findings?

Response:

The SHAP summary plot shows that **num__ApplicantIncome** is the most important feature influencing predictions across all data points, with a wide range of SHAP values indicating strong impact.

Interestingly, this feature has a mixed effect—in many cases, h**igh income (shown in red) pushes predictions toward Class 1 (e.g., approval), but in other instances**, it has little or even negative influence. This variability likely arises from interactions with other features, such as credit history or loan amount.

The color gradient of the dots helps interpret how low (blue) or high (red) feature values relate to the prediction, revealing patterns and exceptions.

Overall, **while some features consistently support predictions in one direction,** others like applicant income show context-dependent behavior, highlighting the complexity of the model's decision-making.

Step 4: Ethical Reflection

As you explored the SHAP plots, you were able to see how individual features contribute to a model's predictions, increasing transparency. However, even with these interpretability tools, there are broader ethical concerns regarding bias and transparency in machine learning systems.

Machine learning models, particularly those used in high-stakes areas like healthcare, finance, and criminal justice, can inadvertently reinforce societal biases present in the training data (e.g., based on race, gender, income). While tools like SHAP provide insights into how models make decisions, they do not prevent biased outcomes.

1. How can explainability tools such as SHAP help identify and address bias in machine learning models?

Answer:

Explainability tools like SHAP enhance transparency by revealing how individual features influence model predictions. Although SHAP doesn't eliminate bias, it plays a crucial role in uncovering it. By examining SHAP values, we can identify when sensitive attributes such as gender, race, or income are having an undue influence on predictions. For example, if a feature like `Gender` consistently contributes significantly to model outputs, it may indicate potential bias.

SHAP also allows for group-level analysis, helping to detect disparities in how features affect different populations. This can highlight unfair treatment or unequal access to outcomes, especially in high-stakes areas like healthcare or finance. Additionally, SHAP supports fairness auditing by enabling comparisons across subgroups, helping practitioners spot patterns that would otherwise be hidden in a black-box model.

While SHAP can help identify sources of bias and guide mitigation strategies—such as re-engineering features or refining training data—it must be used alongside broader ethical practices. SHAP explains what the model is doing, but it does not determine whether the model's use of features is fair or appropriate. Therefore, SHAP should be part of a comprehensive approach to fairness that includes ethical guidelines, bias audits, and careful data governance.

2. How does explainability contribute to transparency in machine learning systems?

Answer:

Explainability tools like SHAP enhance transparency by making complex machine learning models more interpretable, especially for stakeholders such as users, auditors, and regulators. SHAP clarifies the "why" behind each prediction by showing which features influenced the outcome and to what extent—supporting trust, accountability, and informed decision-making in sensitive domains like finance, healthcare, and criminal justice.

It also makes models auditable by providing a shared framework for assessing fairness and ethics, enabling data scientists, ethicists, and compliance teams to evaluate and document model behavior.

This is particularly valuable in regulated industries where organizations must justify decisions, demonstrate that processes are fair, and show that any bias was unintentional.

Moreover, SHAP empowers non-technical stakeholders—such as customers or loan officers—by helping them understand the key drivers behind automated decisions. Ultimately, SHAP supports ethical AI by identifying potential biases, increasing transparency, and enabling meaningful human oversight through auditing, debugging, and fairness-focused interventions.