Report on Customers dataset

Classify2TeX

January 18, 2025

Contents

1 Exploratory Data Analysis

1.1 Non-Null Count, Dtype of features

The table 1 provides information about the dataset, including the number of non-null values and the data types of each feature.

Table 1: Dataset Columns Information

Index	Column	Non-Null Count	Dtype
0	Customer_ID	4000	object
1	Age	4000	int64
2	Gender	4000	object
3	Annual_Income	4000	int64
4	Spending_Score	4000	int64
5	Region	4000	object
6	Marital_Status	4000	object
7	Num_of_Children	4000	int64
8	Employment_Status	4000	object
9	Credit_Score	4000	int64
10	Online_Shopping_Frequency	4000	int64
11	Target	4000	int64

1.2 Descriptive Statistics

The table 2 provides descriptive statistics for the dataset, including the count, mean, standard deviation, minimum, and maximum values.

Table 2: Dataset Descriptive Statistics

Index	Column Name/Statistic	count	mean	std	min	25%	50%	75%	max
0	Age	4000.0	43.63	14.96	18.0	31.0	43.0	57.0	69.0
1	Annual_Income	4000.0	85708.03	37977.69	20076.0	53163.75	85592.5	119030.0	149989.0
2	Spending_Score	4000.0	49.78	29.01	1.0	25.0	49.0	75.0	99.0
3	Num_of_Children	4000.0	1.97	1.4	0.0	1.0	2.0	3.0	4.0
4	Credit_Score	4000.0	575.12	158.68	300.0	438.0	574.0	712.0	849.0
5	Online_Shopping_Frequency	4000.0	9.6	5.78	0.0	5.0	10.0	15.0	19.0
6	Target	4000.0	0.3	0.46	0.0	0.0	0.0	1.0	1.0

1.3 Distribution of features

This section provides a visual representation of the distribution of features in the dataset using histograms (numerical features) and bar charts (categorical features). These visualizations can help in understanding the data.

1.3.1 Histograms of Numerical columns

The histograms below show the distribution of numerical features in the dataset.

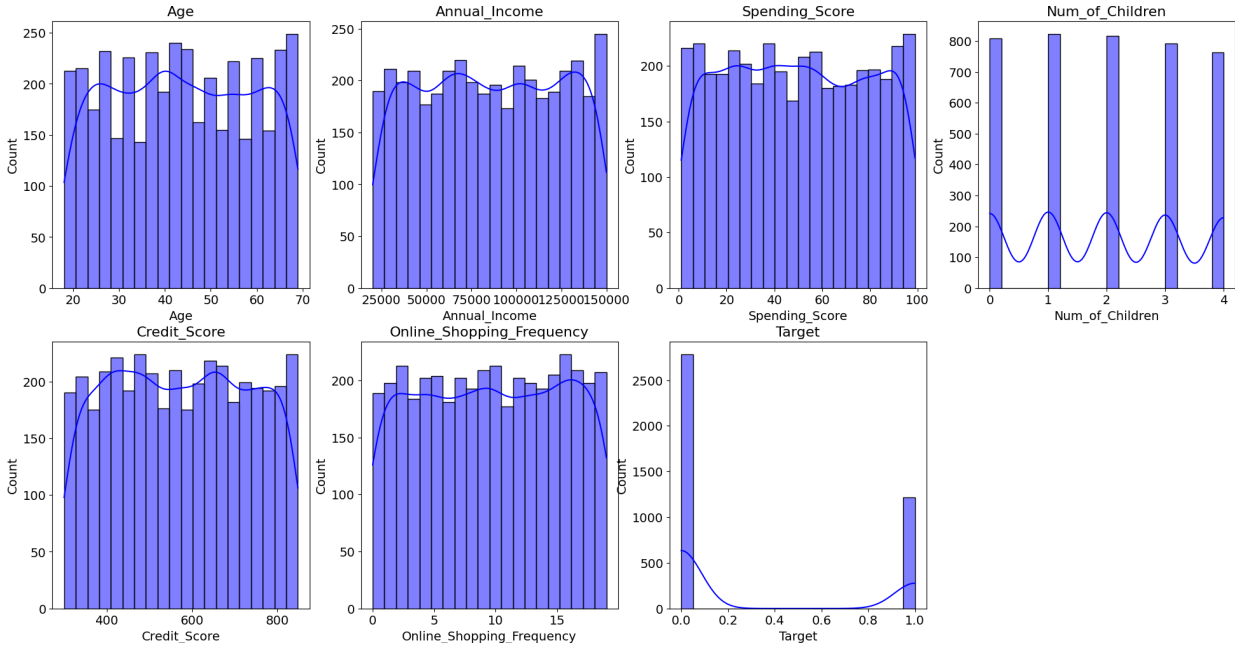

Figure 1: Histograms of Numerical columns

1.3.2 Bar Charts of Categorical columns

The bar charts below show the distribution of categorical features in the dataset.

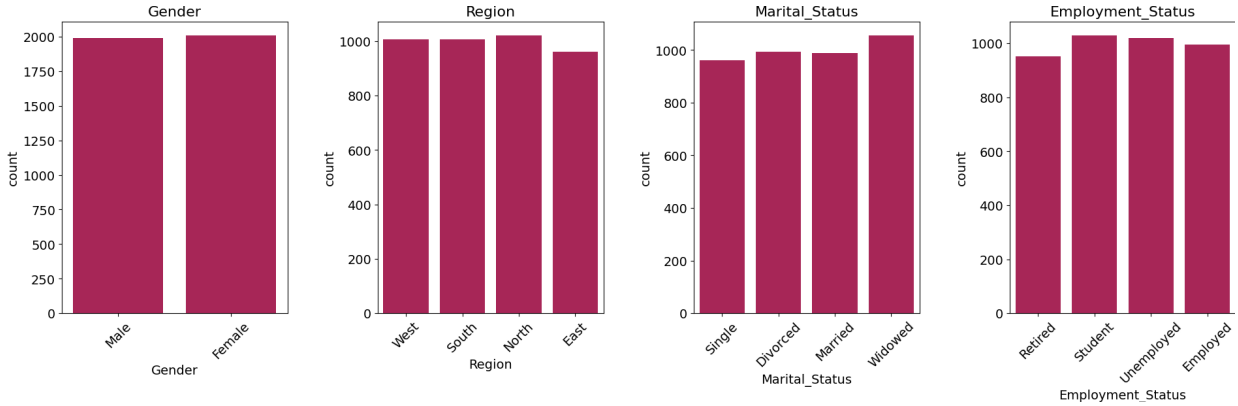

Figure 2: Bar Charts of Categorical columns

2 Evaluation Metrics

2.1 Accuracy

Accuracy is one of the simplest evaluation metrics for classification models. It is defined as the ratio of correctly predicted observations to the total number of observations:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

While accuracy is intuitive and easy to understand, it may not be suitable for imbalanced datasets. For example, in a dataset where 95% of the samples belong to one class, predicting the majority class for every instance would result in high accuracy but poor performance on the minority class.

2.2 F1 Score

The **F1 Score** is the harmonic mean of Precision and Recall, providing a balance between the two. It is particularly useful when dealing with imbalanced datasets. Precision and Recall are defined as follows:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The F1 Score combines these metrics:

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

A high F1 Score indicates a good balance between Precision and Recall, making it a valuable metric in scenarios where false positives and false negatives have significant costs.

2.3 ROC AUC

The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (Recall) against the False Positive Rate at various threshold settings. The **Area Under the Curve (AUC) of the ROC curve** measures the overall ability of the model to distinguish between classes.

$$\text{AUC} = \int_{\text{FPR}=0}^{1} \text{TPR}(\text{FPR})\, d(\text{FPR})$$

Key points about ROC AUC:

- An AUC of 0.5 indicates random guessing.
- An AUC of 1.0 indicates perfect classification.
- It is a threshold-independent metric, providing an aggregate measure of performance across all classification thresholds.

ROC AUC is particularly useful for binary classification tasks and provides insights into the trade-off between sensitivity and specificity.

3 Model Optimization Results

3.1 Optimization Results Tables

The tables below show the hyperparameters and achieved metrics for each model configuration considered during the optimization process. The index of models with default hyperparameters is 0. The next models, indexed from 1, were chosen by Random Search.

Table 3: Random Forest Hyperparameters and achivied metrics

Index	Metric/Hyperp.\ Iteration	0	1	2	3	4	5	6	7
0	f1	0.8716	0.6609	0.8509	0.512	0.5623	0.8897	0.512	0.7825
1	accuracy	0.8716	0.6609	0.8509	0.5122	0.5625	0.8899	0.5122	0.7825
2	roc_auc	0.9192	0.7165	0.9171	0.516	0.5809	0.9186	0.516	0.8611
3	n_estimators	100	50	50	50	200	100	200	200
4	criterion	gini	gini	log_loss	log_loss	gini	entropy	gini	log_loss
5	max_depth	None	20	30	10	10	None	30	10
6	min_samples_split	2	2	2	10	10	2	10	10
7	min_samples_leaf	1	1	1	4	2	2	1	1
8	min_weight_fraction_leaf	0.0	0.01	0.0	0.1	0.05	0.0	0.1	0.0
9	max_features	sqrt	log2	None	None	sqrt	sqrt	None	log2
10	bootstrap	1	1	1	0	1	0	0	1

Table 4: Decision Tree Hyperparameters and achivied metrics

Index	Metric/Hyperp. \ Iteration	0	1	2	3	4	5	6	7
0	f1	0.7689	0.4835	0.6827	0.4835	0.5055	0.5947	0.7176	0.5055
1	accuracy	0.7729	0.4887	0.6853	0.4887	0.5104	0.5955	0.718	0.5104
2	roc_auc	0.7729	0.4851	0.7361	0.4851	0.5127	0.6454	0.7658	0.5127
3	criterion	gini	log_loss	log_loss	gini	gini	entropy	entropy	entropy
4	splitter	best	best	best	best	random	best	random	best
5	max_depth	None	None	40	10	40	10	40	40
6	min_samples_split	2	10	2	10	5	5	5	5
7	min_samples_leaf	1	2	4	4	1	1	1	4
8	max_features	None	None	sqrt	None	None	None	log2	log2
9	class_weight	None	None	None	None	balanced	balanced	balanced	balanced
10	min_impurity_decrease	0.0	0.1	0.0	0.01	0.05	0.0	0.0	0.1

Table 5: XGBoost Hyperparameters and achivied metrics

Index	Metric/Hyperp. \ Iteration	0	1	2	3	4	5	6	7
0	f1	0.8183	0.783	0.6329	0.8337	0.7329	0.6534	0.8183	0.765
1	accuracy	0.8187	0.783	0.6329	0.8339	0.7329	0.6534	0.8188	0.7654
2	roc_auc	0.8852	0.859	0.6877	0.9061	0.8064	0.7163	0.8946	0.8354
3	eval_metric	logloss	logloss	logloss	logloss	logloss	logloss	logloss	logloss
4	n_estimators	100	50	50	100	50	200	200	100
5	max_depth	6	10	6	15	10	6	15	6
6	learning_rate	0.3	0.05	0.05	0.1	0.1	0.01	0.1	0.2
7	subsample	1.0	0.7	0.5	0.9	0.9	0.5	0.7	1.0
8	colsample_bytree	1.0	0.7	0.7	0.7	0.5	0.7	0.9	0.9
9	min_child_weight	1	1	7	3	7	5	5	3
10	gamma	0.0	0.0	0.1	0.2	0.0	0.0	0.1	0.2
11	reg_alpha	0.0	1.0	1.0	0.0	1.0	0.01	0.1	0.0
12	reg_lambda	1.0	1.0	2.0	1.0	1.0	1.0	1.0	1.5

3.2 Boxplots of accuracy, f1, roc_auc

Boxplots of accuracy, F1, and ROC AUC illustrate the distribution and variability of model performance metrics across different configurations of hyperparameters. The plots are located below.

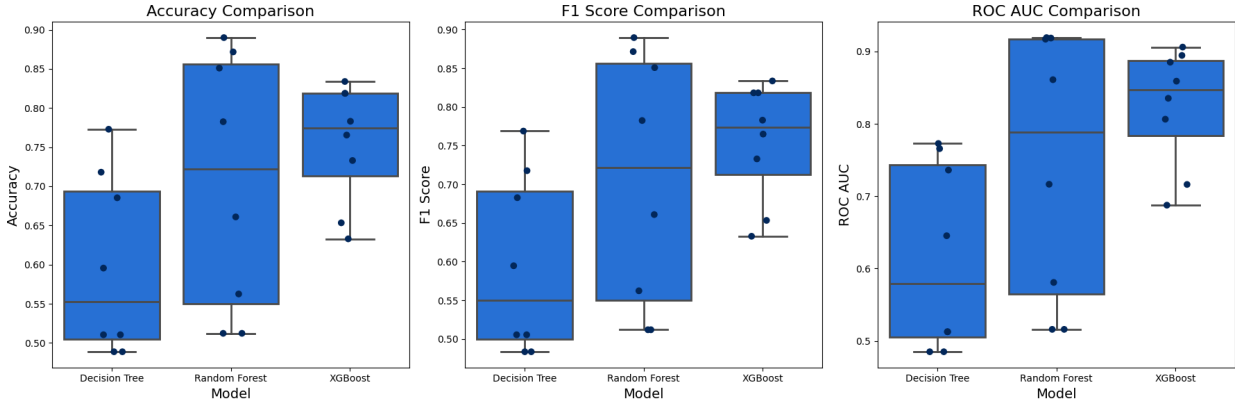

Figure 3: Boxplots of accuracy, f1, roc_auc

3.3 Barplots of maximum values of metrics achievied by model

Barplots of maximum metric values show the highest performance scores for each model type. The plots are located below.

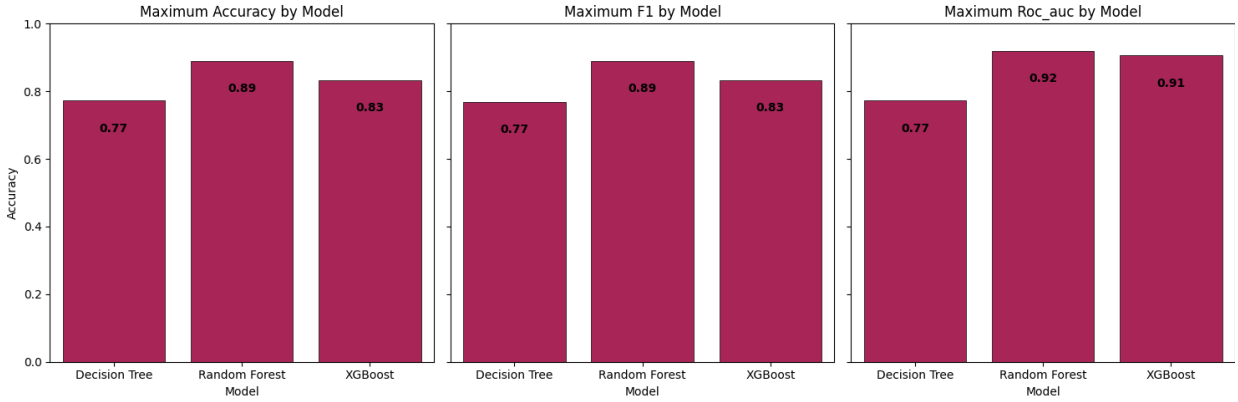

Figure 4: Barplots of maximum values of metrics achievied by model

4 Interpretabilty of the best models

Classify2TeX package defined the best model as the one that achievied the highest value of a metric, chosen by the user, or ROC AUC by default.In this case, the optimization process was aimed at maximizing **ROC AUC.**
Do not forget, that after preprocessing, columns names have changed, because of transformations of categorical features.

4.1 SHAP - what is under the hood?

SHAP (SHapley Additive exPlanations) values are a unified measure of feature importance, grounded in cooperative game theory, that explain the contribution of each feature to the predictions of a machine learning model. By assigning a consistent and fair contribution to each feature, SHAP values offer insights into the underlying decision-making process of the model, both for specific predictions and overall feature importance.

The fundamental principle behind SHAP is that a model's prediction for a given instance can be decomposed into the sum of contributions from its features, along with a baseline value. The baseline typically represents the average model prediction across the dataset when no feature information is provided.

How SHAP Values work

For a specific instance, SHAP calculates how much each feature contributes to the difference between the baseline and the model's prediction. This involves:

1. **Marginal Contributions**: Evaluating how the prediction changes when each feature is added to subsets of other features. For example, if you have features A, B, C, SHAP will compute how the prediction changes when A is added to subsets like $\{\}, \{B\}, \{C\}, \{B, C\}$, etc.
2. **Weighted Averaging Across Subsets**: To compute the SHAP value for a feature, the method takes the average of its marginal contributions across all subsets of features, weighted by the size of the subsets. This ensures fairness in the distribution of contributions.
3. **Baseline Value**: The baseline is a reference point, usually the average model prediction over the dataset. It represents what the model predicts when no features are considered.

For any given data point, SHAP values indicate how much each feature shifts the model's prediction relative to the baseline. A positive SHAP value means the feature increases the prediction, while a negative SHAP value means it decreases the prediction. This decomposition allows for a granular understanding of both the direction and magnitude of each feature's influence on the model's decision.

4.2 The best XGBoost model Explanation

4.2.1 XGBoost model - feature importance using SHAP values

SHAP bar plot provides a concise overview of the importance of individual features in the model's predictions. Each bar represents a feature, with its length corresponding to the mean absolute SHAP value across all samples. This indicates the average magnitude of the feature's contribution to the predictions, regardless of direction.

Features are ranked in descending order of importance, and only the top 15 features are displayed by default for clarity. The bar plot allows quick identification of the most influential features driving the model's behavior and is particularly useful for comparing their relative contributions.

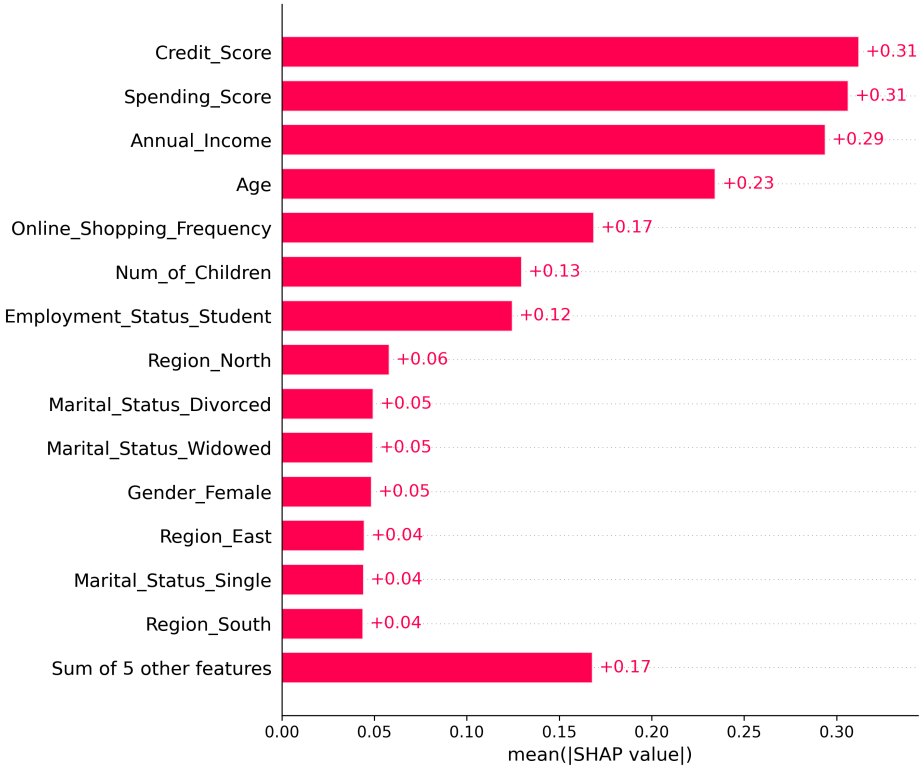

Figure 5: SHAP values for the best XGBoost model

4.2.2 XGBoost model - feature importance gained directly from the model

Feature importance bar chart: visually represents the contributions of individual features to the model's predictions, based on their calculated importance scores. Each bar in the chart corresponds to a feature, and its length indicates the magnitude of that feature's importance in reducing split impurity during the model's training process. Features that contribute more significantly to the model's predictive accuracy are displayed with longer bars, while less influential features have shorter bars.

The chart is horizontally oriented, with feature names listed on the vertical axis and their corresponding importance values on the horizontal axis.

This visualization is especially useful for diagnosing the model's behavior, understanding which features drive its decisions, and identifying variables that have the most impact on predictions.

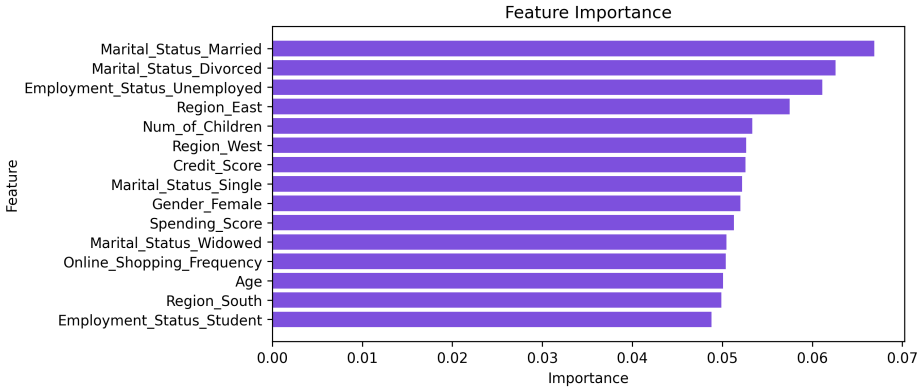

Figure 6: Feature Importance for the best XGBoost model

4.2.3 XGBoost model - violin plot (SHAP) of impact on prediction

SHAP violin plot provides a visual summary of how each feature influences model predictions and the variability of this influence. Features are listed on the vertical axis in descending order of importance, while the horizontal axis shows SHAP values, indicating the magnitude and direction of each feature's contribution to predictions.

The shape of each 'violin' represents the distribution of SHAP values for a feature: wider sections indicate higher density of similar values, while narrower sections show less frequent SHAP values. Positive SHAP values increase the prediction, and negative values decrease it.

Colors correspond to actual feature values, with red typically representing higher values and blue lower ones. The color distribution along the SHAP scale highlights how feature values affect predictions; for instance, if red dominates the positive side, high feature values increase predictions.

By default, only the top 15 features by importance are displayed, keeping the visualization focused and interpretable.

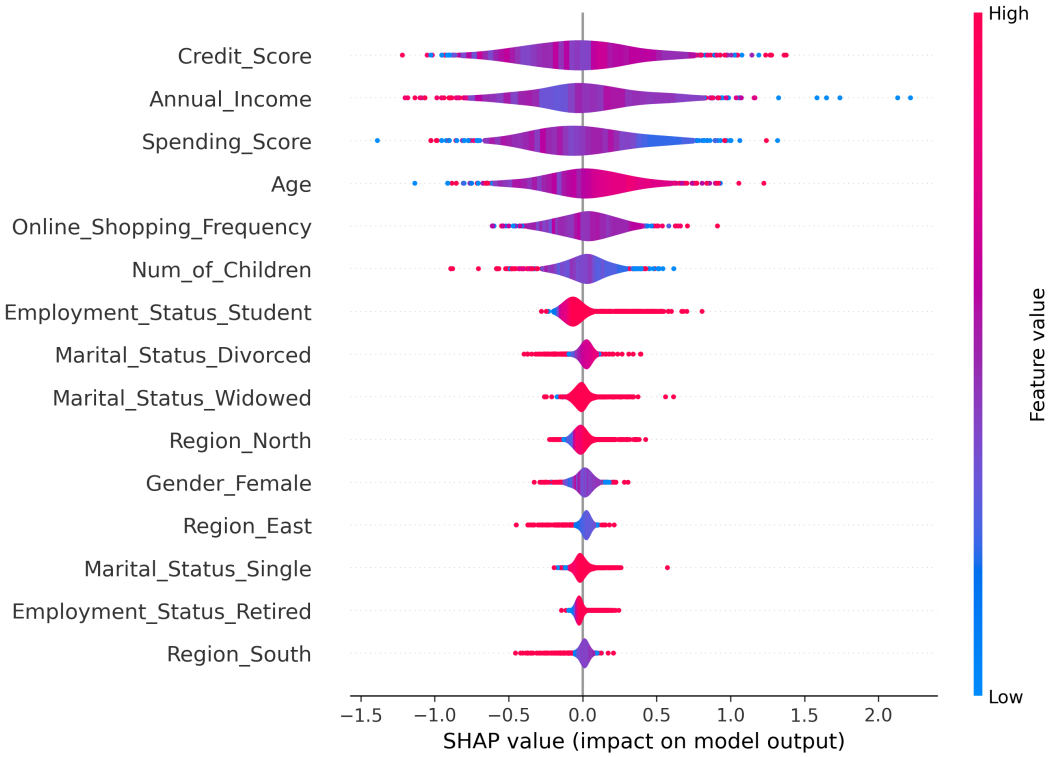

Figure 7: Violin plot (SHAP) of impact on prediction for the best default XGBoost model