# Report on PlacementData dataset

Classify2TeX

January 16, 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 | sl_no | 215 | int64 |
| 1 | gender | 215 | object |
| 2 | ssc_p | 215 | float64 |
| 3 | ssc_b | 215 | object |
| 4 | hsc_p | 215 | float64 |
| 5 | hsc_b | 215 | object |
| 6 | hsc_s | 215 | object |
| 7 | degree_p | 215 | float64 |
| 8 | degree_t | 215 | object |
| 9 | workex | 215 | object |
| 10 | etest_p | 215 | float64 |
| 11 | specialisation | 215 | object |
| 12 | mba_p | 215 | float64 |
| 13 | status | 215 | object |
| 14 | salary | 148 | float64 |

## 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 | sl_no | 215.0 | 108.0 | 62.21 | 1.0 | 54.5 | 108.0 | 161.5 | 215.0 |
| 1 | ssc_p | 215.0 | 67.3 | 10.83 | 40.89 | 60.6 | 67.0 | 75.7 | 89.4 |
| 2 | hsc_p | 215.0 | 66.33 | 10.9 | 37.0 | 60.9 | 65.0 | 73.0 | 97.7 |
| 3 | degree_p | 215.0 | 66.37 | 7.36 | 50.0 | 61.0 | 66.0 | 72.0 | 91.0 |
| 4 | etest_p | 215.0 | 72.1 | 13.28 | 50.0 | 60.0 | 71.0 | 83.5 | 98.0 |
| 5 | mba_p | 215.0 | 62.28 | 5.83 | 51.21 | 57.95 | 62.0 | 66.25 | 77.89 |
| 6 | salary | 148.0 | 288655.41 | 93457.45 | 200000.0 | 240000.0 | 265000.0 | 300000.0 | 940000.0 |

## 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.

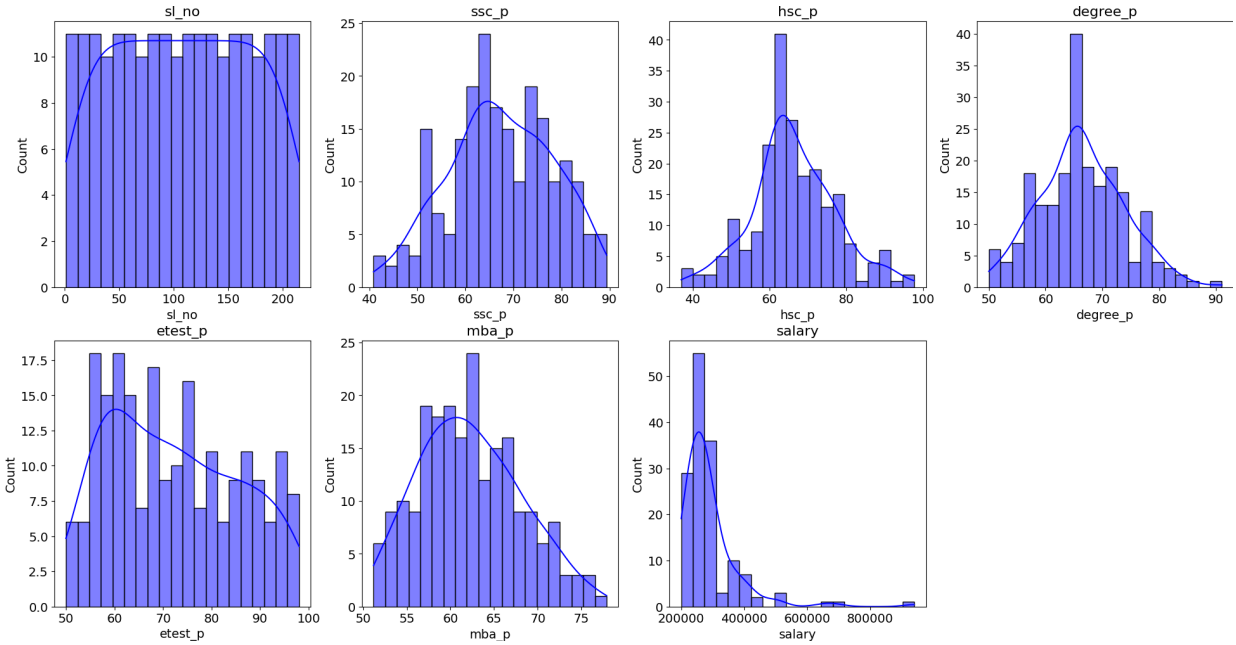


Figure 1: Histograms of Numerical columns

### 1.3.2 Bar Charts of Categorical columns

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

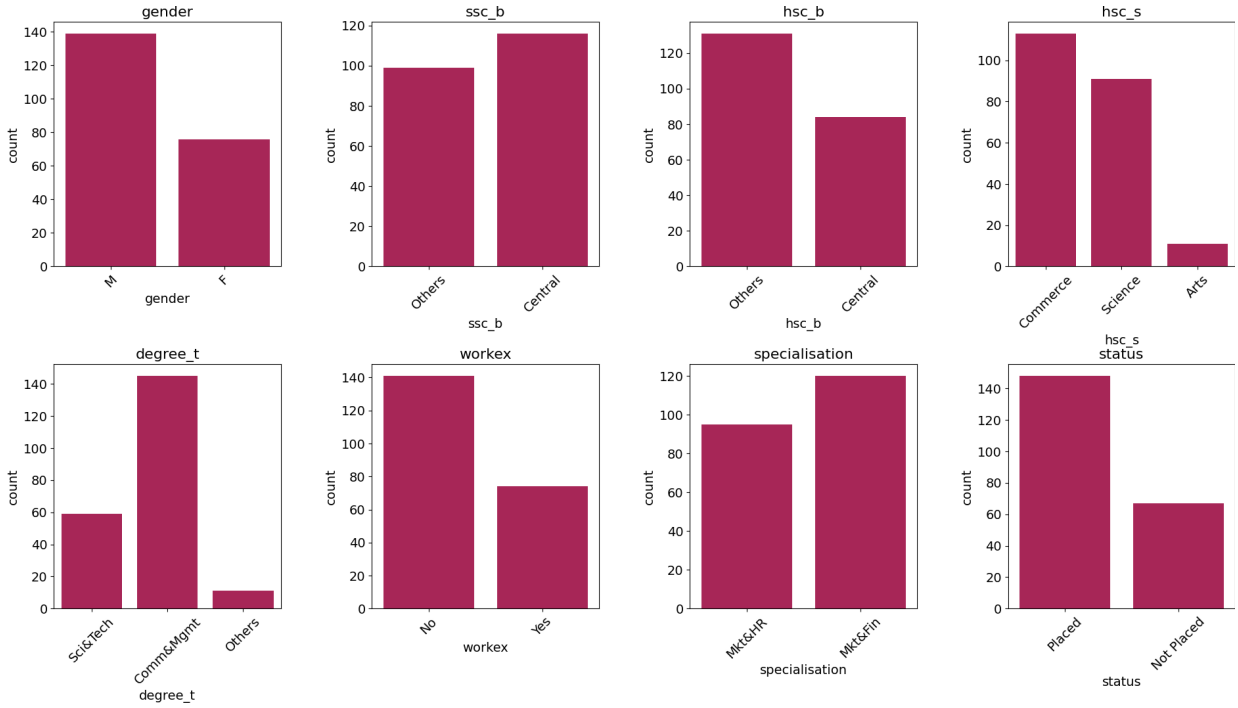


Figure 2: Bar Charts of Categorical columns

# 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

Table 3: Random Forest Hyperparameters and achivied metrics

| Index | Metric/Hyperp.\ Iteration | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | f1 | 0.9499 | 0.9595 | 0.9831 | 0.9763 | 0.9797 | 0.9797 | 0.9561 | 0.9899 | 0.8943 |
| 1 | accuracy | 0.95 | 0.9595 | 0.9831 | 0.9764 | 0.9797 | 0.9797 | 0.9561 | 0.9899 | 0.8953 |
| 2 | roc_auc | 0.9922 | 0.994 | 0.9985 | 0.9983 | 0.9983 | 0.9985 | 0.9947 | 0.993 | 0.9692 |
| 3 | n_estimators | 100 | 200 | 50 | 50 | 50 | 200 | 50 | 500 | 500 |
| 4 | criterion | gini | entropy | entropy | gini | log_loss | entropy | entropy | entropy | entropy |
| 5 | max_depth | None | 30 | None | 20 | 30 | 20 | 20 | 10 | 10 |
| 6 | min_samples_split | 2 | 5 | 2 | 10 | 2 | 10 | 5 | 5 | 5 |
| 7 | min_samples_leaf | 1 | 1 | 2 | 1 | 4 | 2 | 4 | 1 | 2 |
| 8 | min_weight_fraction_leaf | 0.0 | 0.05 | 0.0 | 0.01 | 0.0 | 0.0 | 0.05 | 0.0 | 0.1 |
| 9 | max_features | sqrt | sqrt | sqrt | log2 | None | sqrt | log2 | None | None |
| 10 | bootstrap | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 |

Table 4: Decision Tree Hyperparameters and achivied metrics

| Index | Metric/Hyperp. \ Iteration | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | f1 | 0.9333 | 0.8782 | 0.9188 | 0.973 | 0.6318 | 0.8682 | 0.8176 | 0.4607 |
| 1 | accuracy | 0.9333 | 0.8784 | 0.9189 | 0.973 | 0.6351 | 0.8682 | 0.8176 | 0.4662 |
| 2 | roc_auc | 0.9333 | 0.9355 | 0.9487 | 0.9885 | 0.71 | 0.9288 | 0.8987 | 0.4596 |
| 3 | criterion | gini | entropy | gini | log_loss | gini | entropy | entropy | entropy |
| 4 | splitter | best | random | random | best | random | random | random | random |
| 5 | max_depth | None | 30 | None | 10 | 10 | 20 | 40 | 10 |
| 6 | min_samples_split | 2 | 5 | 2 | 5 | 10 | 2 | 10 | 5 |
| 7 | min_samples_leaf | 1 | 1 | 2 | 4 | 1 | 2 | 4 | 1 |
| 8 | max_features | None | sqrt | None | None | None | None | log2 | log2 |
| 9 | class_weight | None | balanced | balanced | None | None | None | balanced | None |
| 10 | min_impurity_decrease | 0.0 | 0.0 | 0.0 | 0.01 | 0.05 | 0.05 | 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 | 0.9333 | 0.9392 | 0.9628 | 0.9899 | 0.9459 | 0.9797 | 0.9763 | 0.9797 |
| 1 | accuracy | 0.9333 | 0.9392 | 0.9628 | 0.9899 | 0.9459 | 0.9797 | 0.9764 | 0.9797 |
| 2 | roc_auc | 0.9967 | 0.9816 | 0.9963 | 0.9999 | 0.9865 | 0.999 | 0.9984 | 0.9996 |
| 3 | eval_metric | logloss | logloss | logloss | logloss | logloss | logloss | logloss | logloss |
| 4 | n_estimators | 100 | 500 | 50 | 500 | 200 | 500 | 100 | 200 |
| 5 | max_depth | 6 | 15 | 10 | 3 | 3 | 10 | 15 | 10 |
| 6 | learning_rate | 0.3 | 0.1 | 0.01 | 0.2 | 0.2 | 0.05 | 0.2 | 0.1 |
| 7 | subsample | 1.0 | 0.5 | 0.9 | 0.7 | 0.5 | 0.9 | 1.0 | 0.7 |
| 8 | colsample_bytree | 1.0 | 0.9 | 0.5 | 0.5 | 0.5 | 0.7 | 0.7 | 0.7 |
| 9 | min_child_weight | 1 | 5 | 1 | 1 | 5 | 1 | 7 | 1 |
| 10 | gamma | 0.0 | 0.0 | 0.1 | 0.1 | 0.2 | 0.1 | 0.0 | 0.2 |
| 11 | reg_alpha | 0.0 | 0.0 | 0.1 | 0.1 | 1.0 | 1.0 | 1.0 | 0.0 |
| 12 | reg_lambda | 1.0 | 5.0 | 1.5 | 2.0 | 2.0 | 5.0 | 1.0 | 5.0 |

## 3.2 Boxplots of accuracy, f1, roc_auc

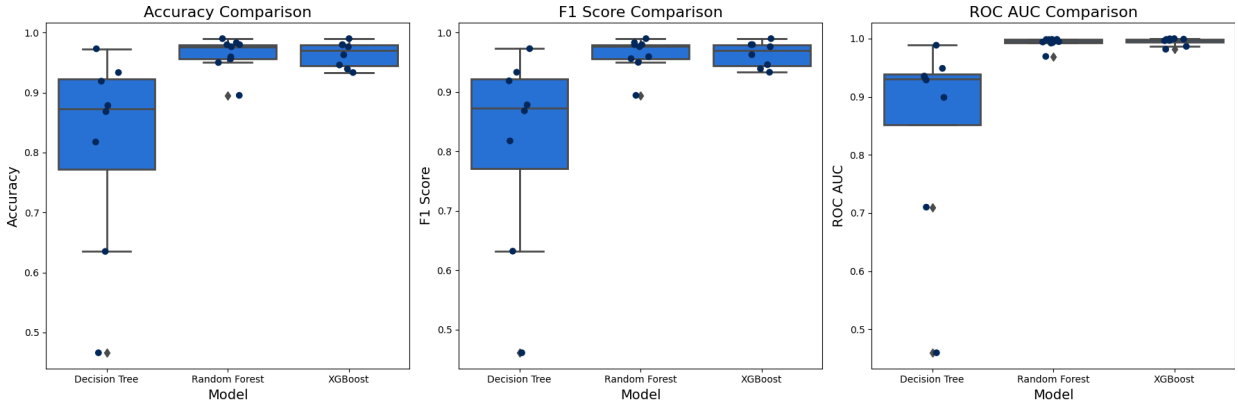


Figure 3: Boxplots of accuracy, f1, roc_auc

## 3.3 Barplots of maximum values of metrics achievied by model

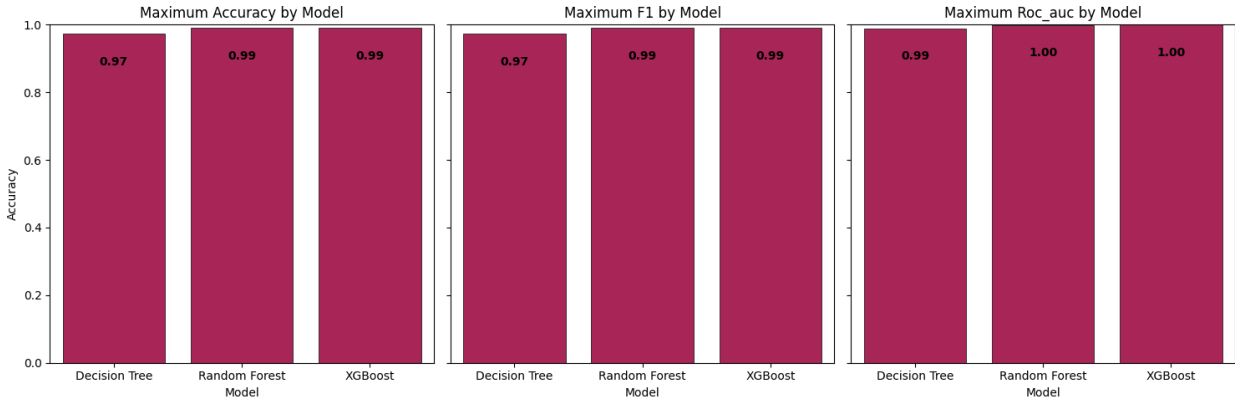


Figure 4: Barplots of maximum values of metrics achievied by model

# 4 Interpretabilty of the best models

Auto2class 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 The best XGBoost model Explanation

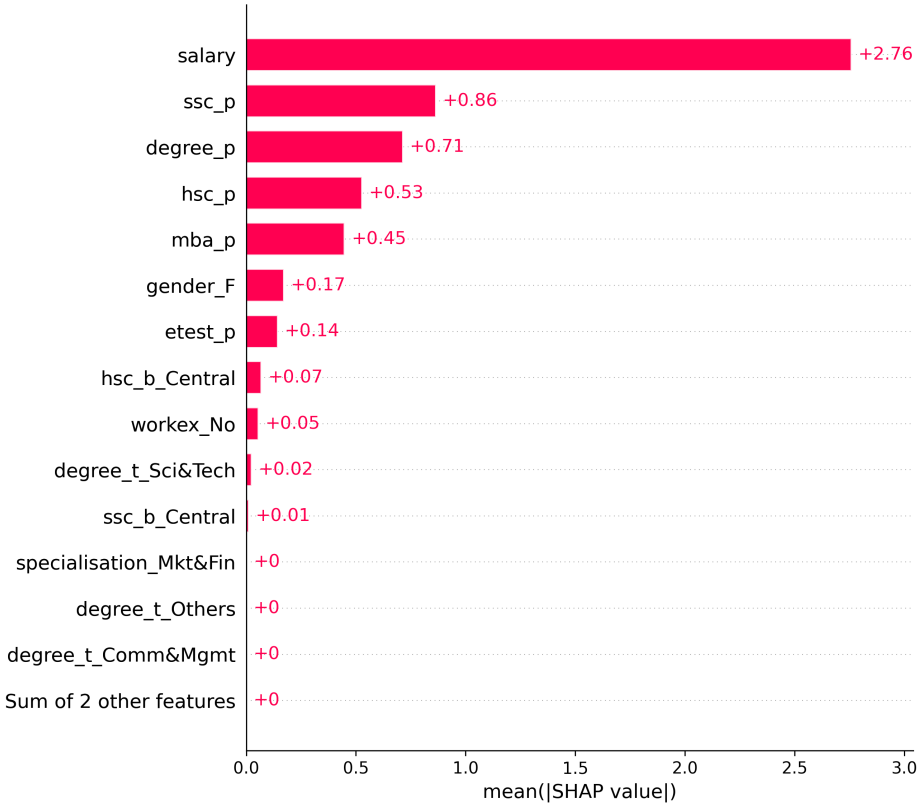


Figure 5: SHAP values for the best XGBoost model

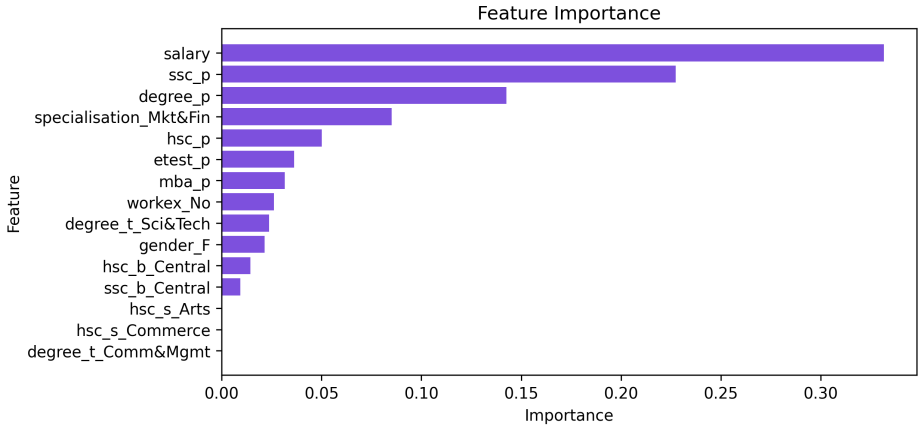


Figure 6: Feature Importance for the best XGBoost model

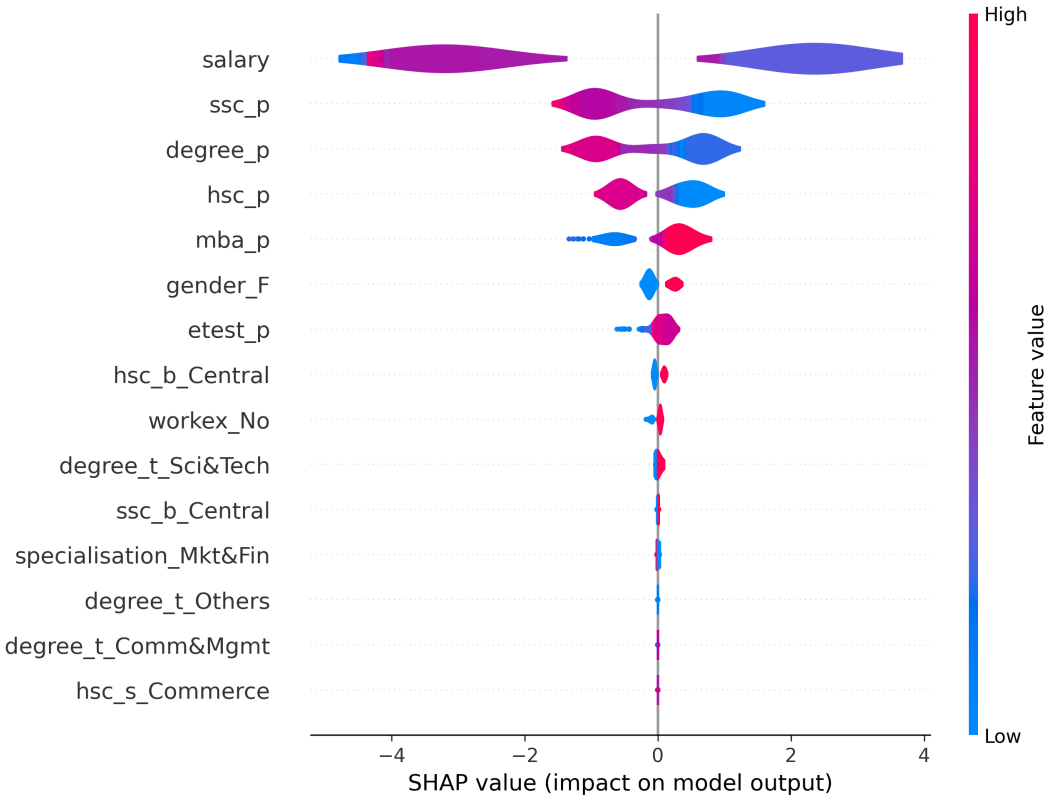


Figure 7: Violin plot (SHAP) of impact on prediction for the best default XGBoost model