Report on PlacementData dataset

Auto2Class

January 13, 2025

Contents

1 Exploratory Data Analysis

1.1 Non-Null Count, Dtype of features

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

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

1.3.1 Histograms of Numerical columns

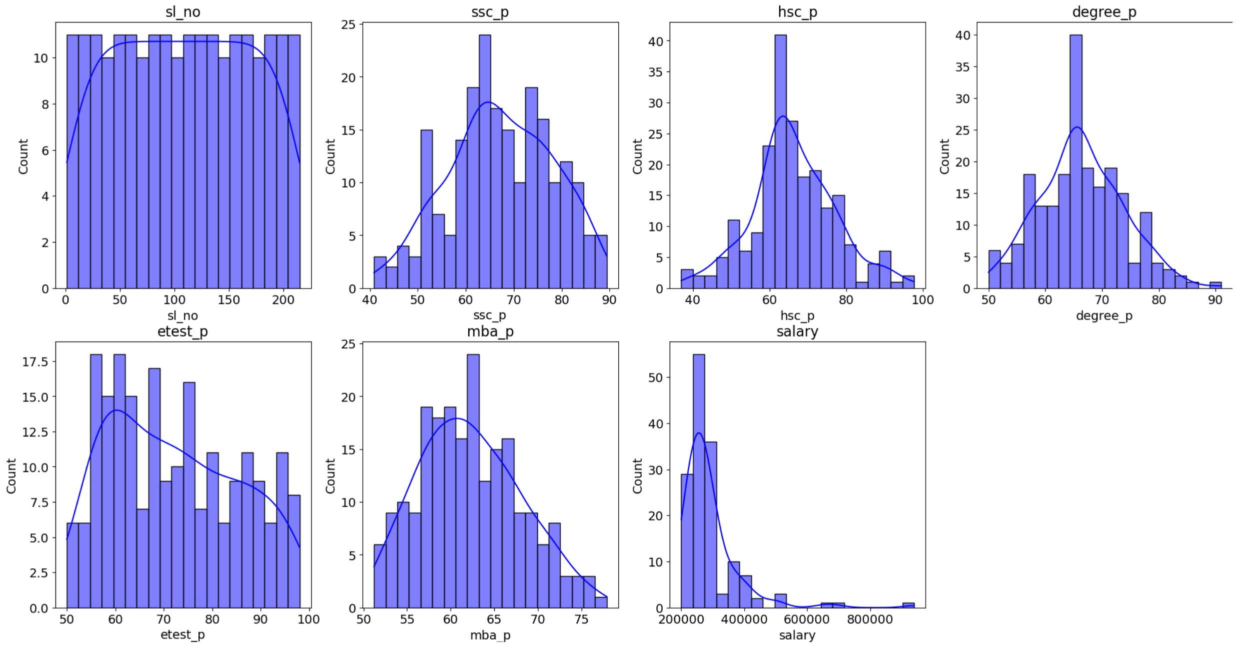

Figure 1: Histograms of Numerical columns

1.3.2 Bar Charts of Categorical columns

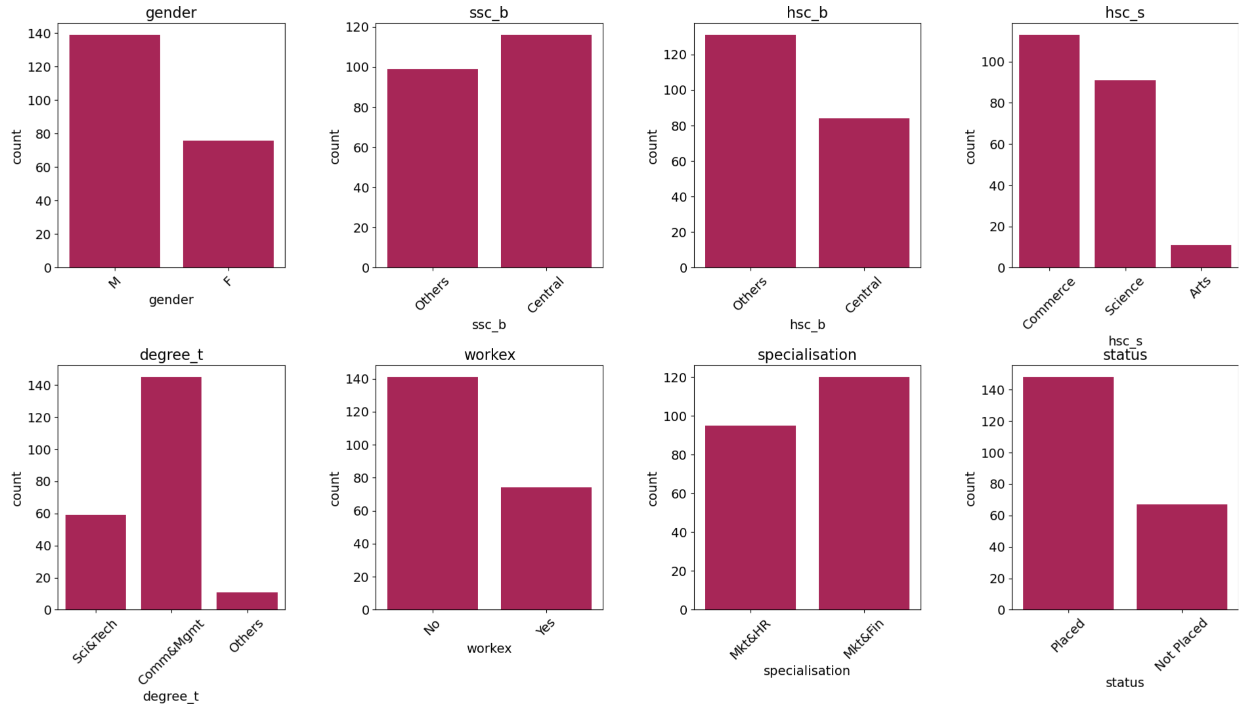

Figure 2: Bar Charts of Categorical columns

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
0	f1	1.0	0.973	0.9797	0.9561	0.9527	0.9763
1	accuracy	1.0	0.973	0.9797	0.9561	0.9527	0.9764
2	roc_auc	1.0	0.9961	0.9986	0.9803	0.9932	0.9986
3	n_estimators	100	50	50	50	200	100
4	criterion	gini	gini	log_loss	log_loss	gini	entropy
5	max_depth	None	20	30	10	10	None
6	min_samples_split	2	2	2	10	10	2
7	min_samples_leaf	1	1	1	4	2	2
8	min_weight_fraction_leaf	0.0	0.01	0.0	0.1	0.05	0.0
9	max_features	sqrt	log2	None	None	sqrt	sqrt
10	bootstrap	1	1	1	0	1	0

Table 4: Decision Tree Hyperparameters and achivied metrics

Index	Metric/Hyperp. \ Iteration	0	1	2	3	4	5
0	f1	0.9833	0.9493	0.8546	0.973	0.7392	0.9831
1	accuracy	0.9833	0.9493	0.8547	0.973	0.7399	0.9831
2	roc_auc	0.9833	0.9681	0.9144	0.9818	0.8162	0.9831
3	criterion	gini	log_loss	log_loss	gini	gini	entropy
4	splitter	best	best	best	best	random	best
5	max_depth	None	None	40	10	40	10
6	min_samples_split	2	10	2	10	5	5
7	min_samples_leaf	1	2	4	4	1	1
8	max_features	None	None	sqrt	None	None	None
9	class_weight	None	None	None	None	balanced	balanced
10	min_impurity_decrease	0.0	0.1	0.0	0.01	0.05	0.0

Table 5: XGBoost Hyperparameters and achivied metrics

Index	Metric/Hyperp. \ Iteration	0	1
0	f1	1.0	0.9797
1	accuracy	1.0	0.9797
2	roc_auc	1.0	0.9969
3	eval_metric	logloss	logloss
4	n_estimators	100	50
5	max_depth	6	10
6	learning_rate	0.3	0.05
7	subsample	1.0	0.7
8	colsample_bytree	1.0	0.7
9	min_child_weight	1	1
10	gamma	0	0
11	reg_alpha	0	1
12	reg_lambda	1	1

3.2 Boxplots of accuracy, f1, roc_auc

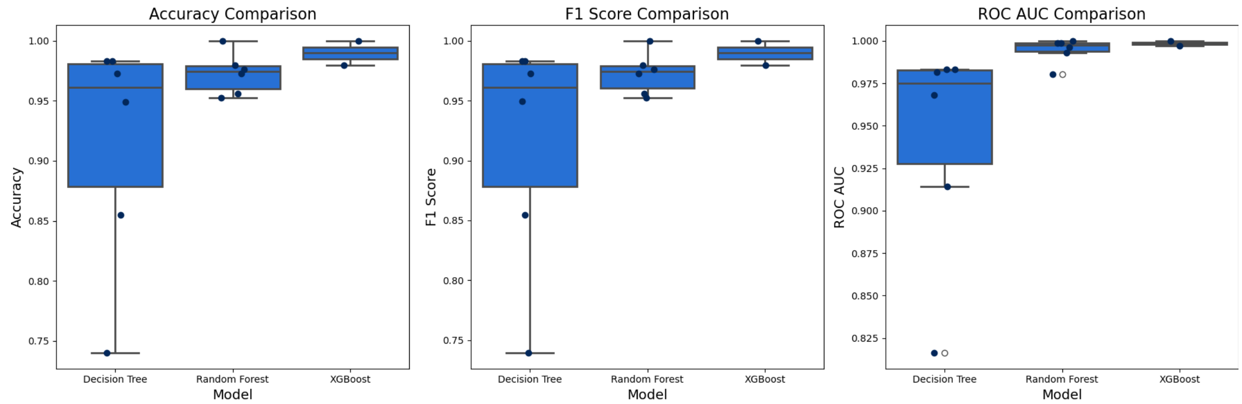

Figure 3: Boxplots of accuracy, f1, roc_auc

3.3 Barplots of maximum values of metrics achievied by model

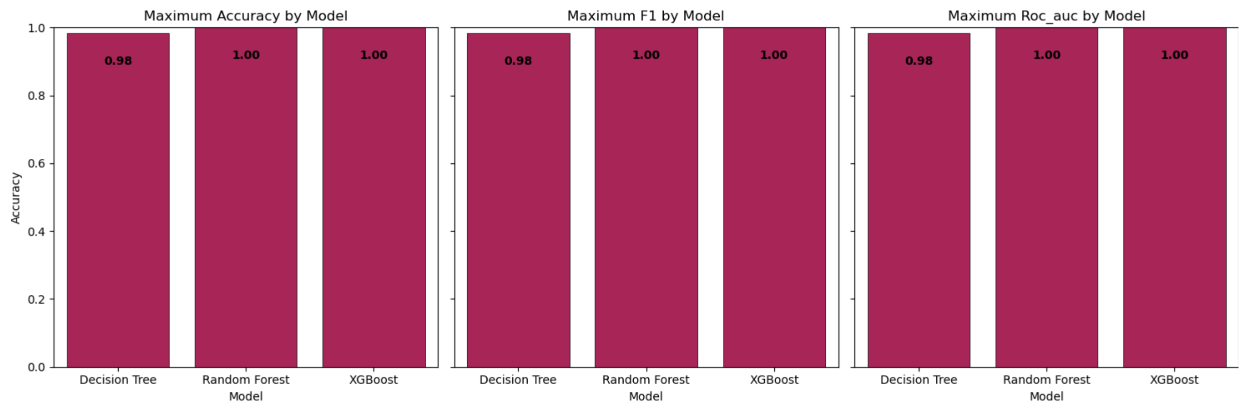

Figure 4: Barplots of maximum values of metrics achievied by model

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 **F1 Score.**
Do not forget, that after preprocessing, columns names have changed, because of transformations of categorical features.

4.1 The best XGBoost model Explanation

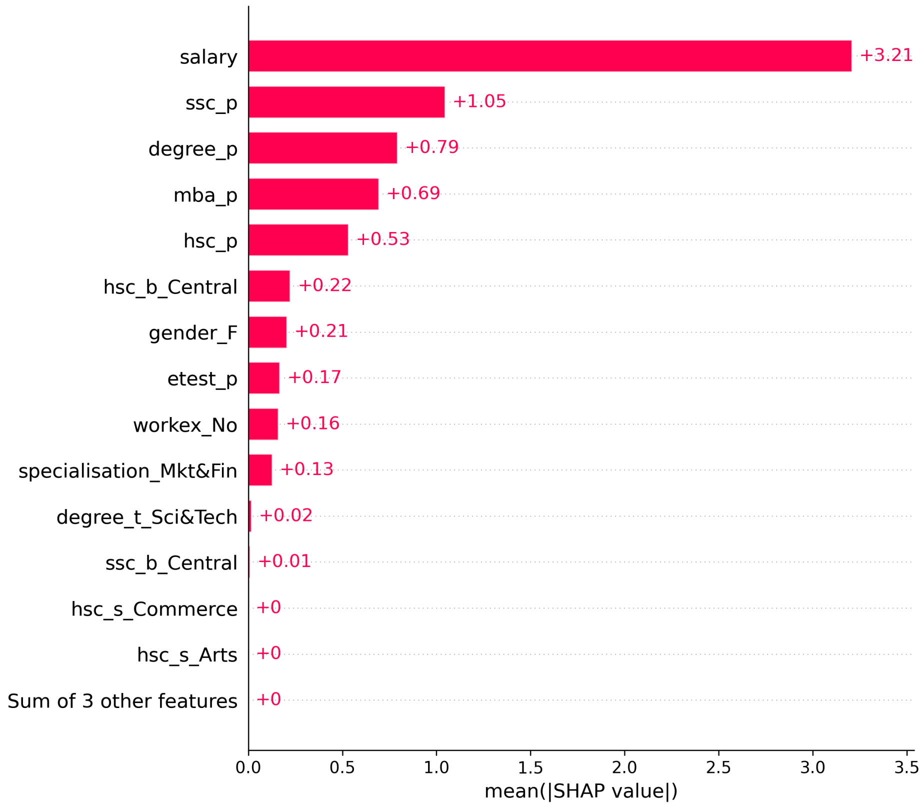

Figure 5: SHAP values for the best XGBoost model

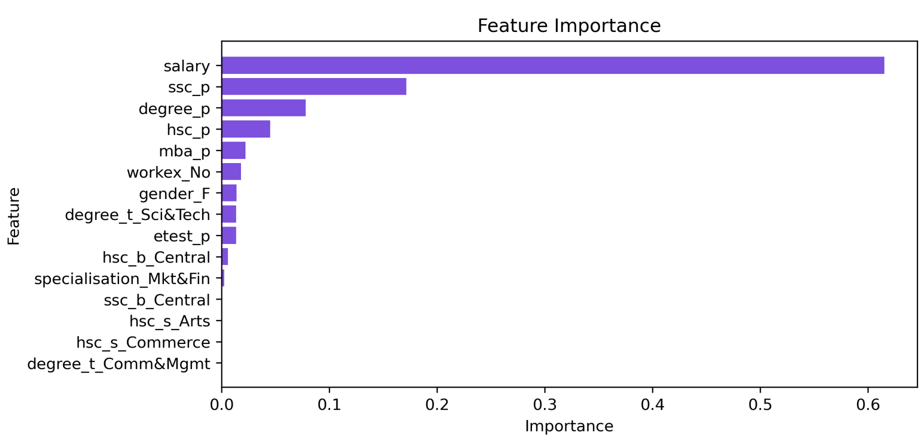

Figure 6: Feature Importance for the best XGBoost model

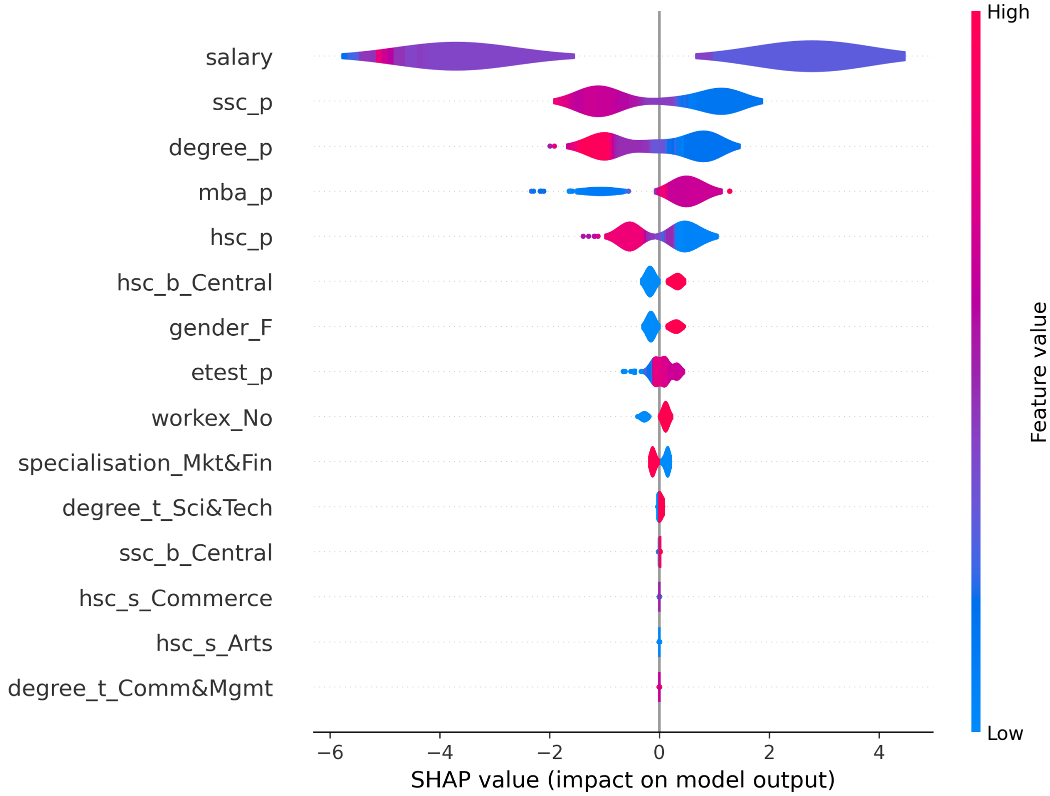

Figure 7: Violin plot (SHAP) of impact on prediction for the best default XGBoost model