Report on Titanic 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	PassengerId	891	int64
1	Survived	891	int64
2	Pclass	891	int64
3	Name	891	object
4	Sex	891	object
5	Age	714	float64
6	SibSp	891	int64
7	Parch	891	int64
8	Ticket	891	object
9	Fare	891	float64
10	Cabin	204	object
11	Embarked	889	object

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	PassengerId	891.0	446.0	257.35	1.0	223.5	446.0	668.5	891.0
1	Survived	891.0	0.38	0.49	0.0	0.0	0.0	1.0	1.0
2	Pclass	891.0	2.31	0.84	1.0	2.0	3.0	3.0	3.0
3	Age	714.0	29.7	14.53	0.42	20.12	28.0	38.0	80.0
4	SibSp	891.0	0.52	1.1	0.0	0.0	0.0	1.0	8.0
5	Parch	891.0	0.38	0.81	0.0	0.0	0.0	0.0	6.0
6	Fare	891.0	32.2	49.69	0.0	7.91	14.45	31.0	512.33

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.

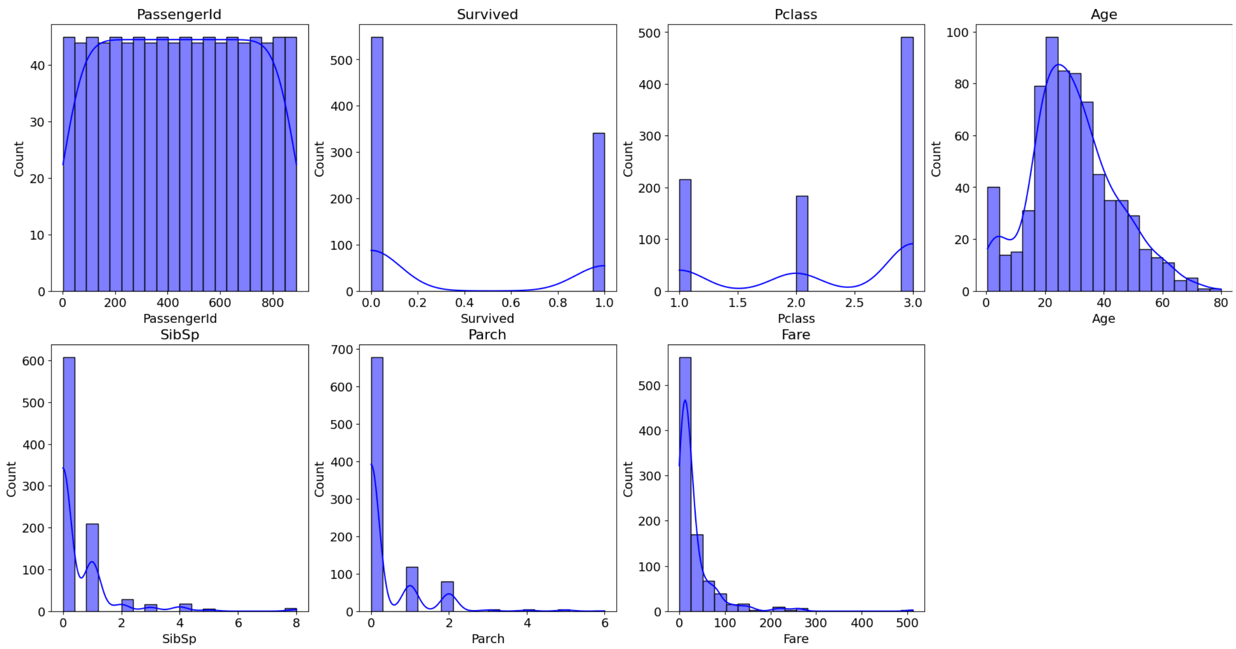

Figure 1: Histograms of Numerical columns

1.3.2 Bar Charts of Categorical columns

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

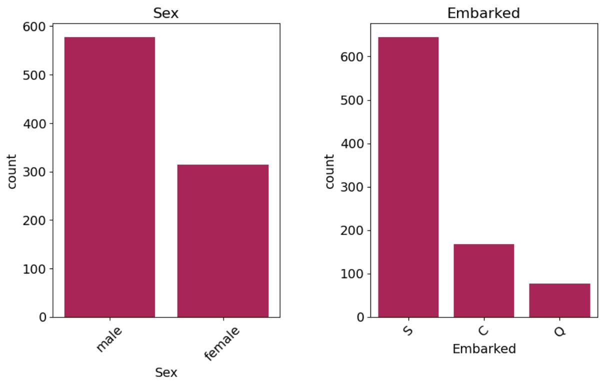

Figure 2: Bar Charts of Categorical columns

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	0.803	0.8256	0.8158	0.7647	0.7992	0.8215	0.7647	0.8343	0.7961
1	accuracy	0.8045	0.8294	0.8171	0.7767	0.807	0.8238	0.7767	0.8373	0.8047
2	roc_auc	0.8292	0.872	0.8612	0.8322	0.8545	0.8652	0.8344	0.8697	0.8537
3	n_estimators	100	50	50	50	200	100	200	200	200
4	criterion	gini	gini	log_loss	log_loss	gini	entropy	gini	log_loss	gini
5	max_depth	None	20	30	10	10	None	30	10	10
6	min_samples_split	2	2	2	10	10	2	10	10	2
7	min_samples_leaf	1	1	1	4	2	2	1	1	2
8	min_weight_fraction_leaf	0.0	0.01	0.0	0.1	0.05	0.0	0.1	0.0	0.05
9	max_features	sqrt	log2	None	None	sqrt	sqrt	None	log2	sqrt
10	bootstrap	1	1	1	0	1	0	0	1	0

Table 4: Decision Tree Hyperparameters and achivied metrics

Index	Metric/Hyperp. \ Iteration	0	1	2	3	4	5	6	7
0	f1	0.8195	0.7849	0.7764	0.8059	0.7849	0.8006	0.781	0.3907
1	accuracy	0.8212	0.7868	0.7778	0.8103	0.7868	0.8013	0.78	0.4355
2	roc_auc	0.7949	0.7435	0.8242	0.8434	0.7435	0.8128	0.8143	0.5081
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	0.7994	0.8273	0.7931	0.8319	0.8175	0.8025	0.8375	0.8335
1	accuracy	0.7989	0.8316	0.8036	0.8339	0.8215	0.8114	0.8395	0.8361
2	roc_auc	0.809	0.8764	0.8506	0.8802	0.8665	0.8593	0.881	0.8787
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

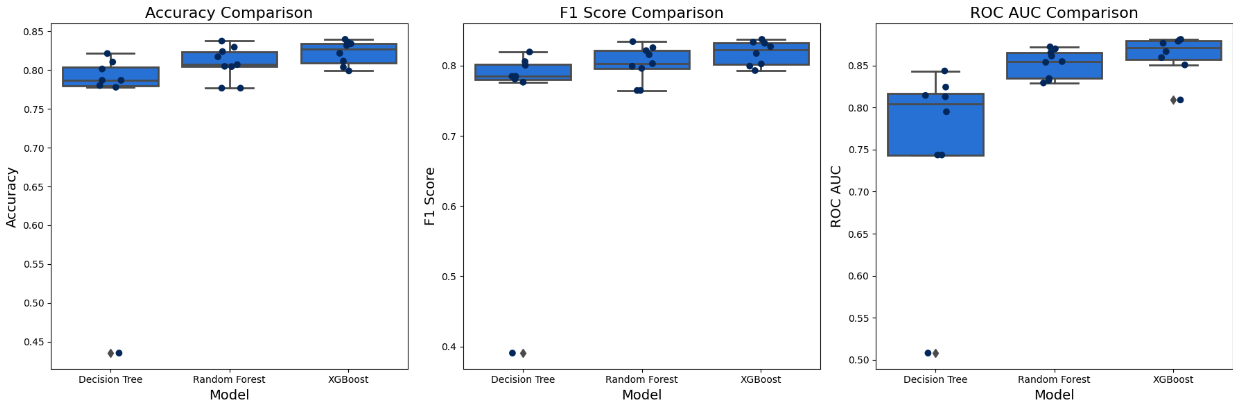

Figure 3: Boxplots of accuracy, f1, roc_auc

3.3 Barplots of maximum values of metrics achievied by model

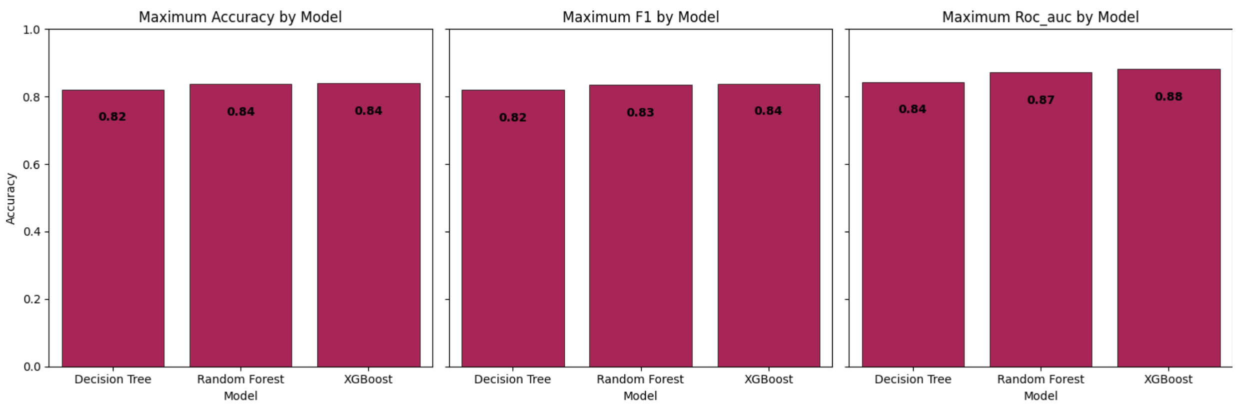

Figure 4: Barplots of maximum values of metrics achievied by model

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

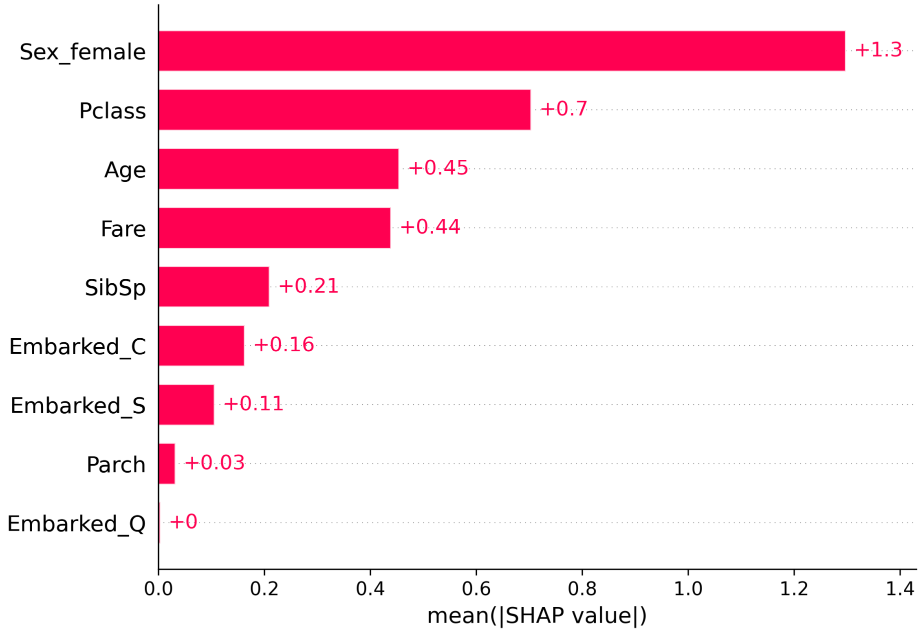

Figure 5: SHAP values for the best XGBoost model

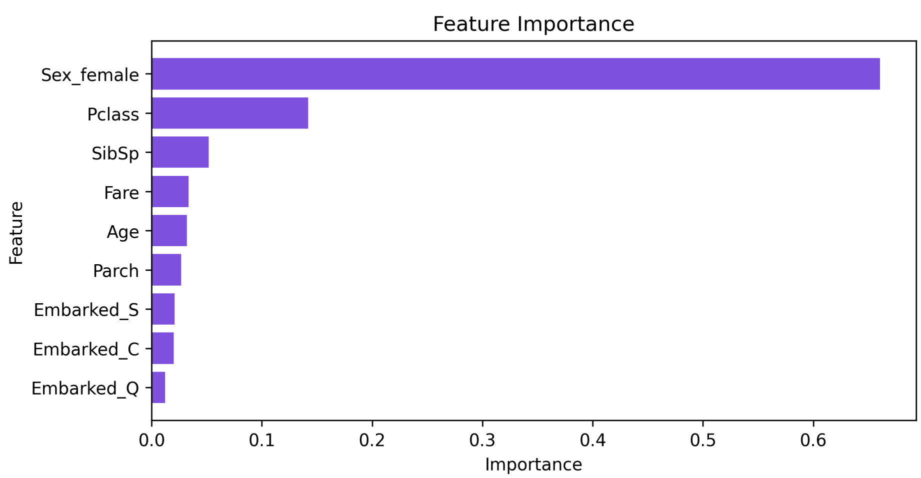

Figure 6: Feature Importance for the best XGBoost model

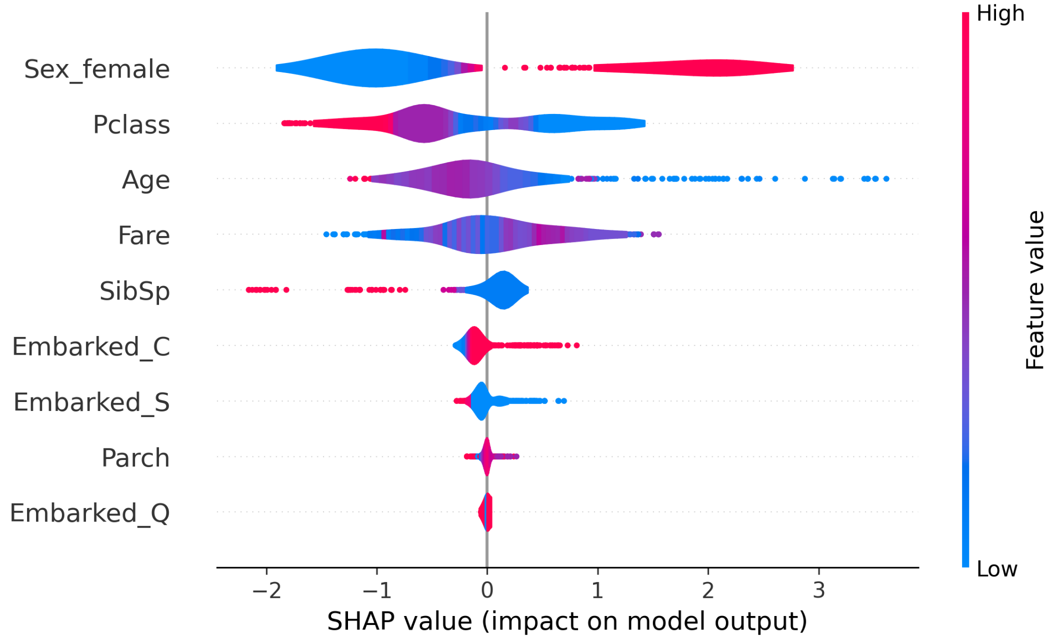

Figure 7: Violin plot (SHAP) of impact on prediction for the best default XGBoost model