Introduction

This notebook is designed to build a model to predict hotel booking cancellations. For this purpose, the dataset `booking.csv` was used. It includes columns describing booking details and a column indicating whether the booking was canceled.

Data Cleaning and Preprocessing

The initial data required some cleaning. Firstly, the date column contained dates in various formats. Additionally, some dates were invalid (e.g., 29th of February in non-leap years).

To prepare the data for modeling, sine and cosine transformations were applied to the `day`, `month`, and `day of the week` features to appropriately capture cyclical relationships. One-hot encoding was applied to categorical features, except when using the CatBoost classifier.

Model Selection

Hyperparameter tuning was performed for the following models: Logistic Regression, XGBoost, CatBoost, and a Feedforward Neural Network. Models were evaluated based on the ROC AUC score.

Logistic Regression

For Logistic Regression:

- Cross-validation and grid search were employed for hyperparameter tuning of the regularization parameter C.
- Standard scaling was applied as part of data preprocessing.
- L2 regularization was used.

The best model was achieved with C=1.0023, resulting in a ROC AUC score of **0.8667**.

XGBoost

For the XGBoost model:

- Cross-validation and grid search were used to tune `learning_rate`, `max_depth`, and `n_estimators`.

The best model was achieved with:

- `learning_rate = 0.1033`
- `max_depth = 11`
- `n_estimators = 180`

This resulted in a ROC AUC score of **0.9594**.

CatBoost

For the CatBoost model:

- Cross-validation and grid search were performed to tune `learning_rate`, `max_depth`, and `n_estimators`.

The best model was achieved with:

- `learning_rate = 0.17`
- `max_depth = 9`
- `n_estimators = 190`

This resulted in a ROC AUC score of **0.9554**.

Feedforward Neural Network

For the Feedforward Neural Network:

- The architecture included 3 hidden layers with ReLU activation functions and a decreasing number of units.
- Cross-validation and random search were used to tune the hyperparameters `learning_rate`, `dropout_rate`, and the number of units in the input layer.
- Early stopping was employed during training.

The best model was achieved with:

- `learning_rate = 0.001`
- `dropout_rate = 0.3`
- `number of units in the input layer = 128`

This resulted in a ROC AUC score of **0.8702**.

Conclusion

The XGBoost model achieved the best performance in predicting booking cancellations, with a ROC AUC score of **0.9594**. This model was saved in JSON format for future use.