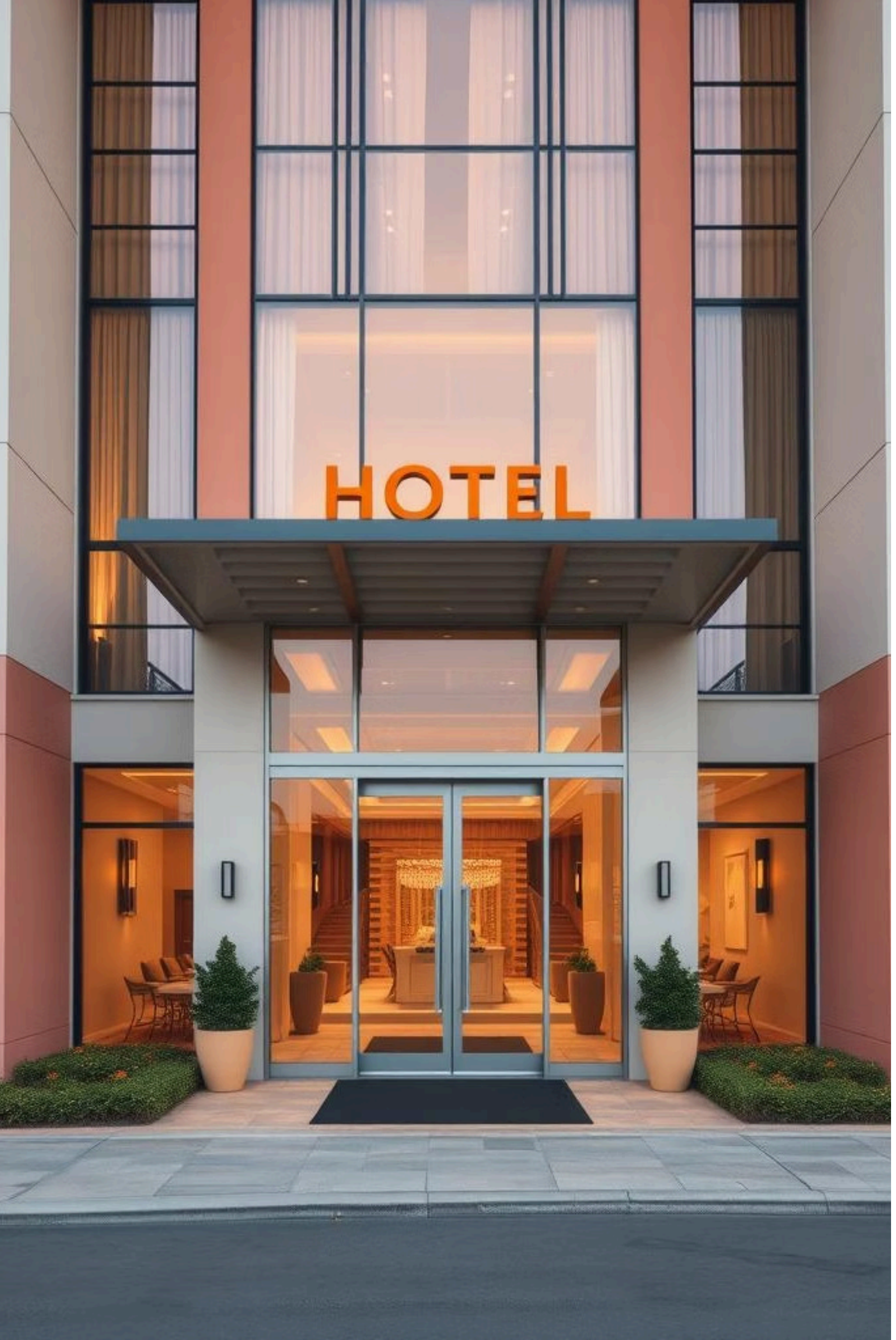


Data Preprocessing for Binary Classification: Booking Status

This presentation outlines the essential steps in preparing data for a binary classification task, specifically predicting booking status. Effective data preprocessing is crucial for building robust and accurate machine learning models. We'll cover everything from cleaning and handling outliers to feature engineering and model evaluation.

Identify Missing Values

Use methods like `.isnull().sum()` or `.info()` in pandas to locate nulls.

Verify & Convert Data Types

Convert *"date of reservation"* column into datetime format.

Remove Duplicates

Remove duplicated records from the whole dataset using `.drop_duplicates()` to ensure consistency.

Imputation Strategies

Decide on handling missing data: "drop rows" as the number is few comparing to the whole data.

Why Handle Outliers?

Outliers can significantly distort model training, leading to biased results and reduced predictive performance.

Z-Score Method for Detection

Outliers are typically defined as data points with an absolute z-score greater than **3** standard deviations from the mean.

Handling Strategies

filling with the median value.

Feature Selection

Eliminate irrelevant or redundant features using correlation analysis and also solving the multicollinearity problem.

Feature Extraction

Create new, more informative features from existing ones, like **"family members"** from **"number of adults"** and **"number of children"**

and also **"total nights"** from **"number of week nights"** and **"number of weekend nights"**.

Transforming Categorical Data

Machine learning models typically require numerical input. Therefore, categorical variables must be converted. The choice of encoding depends on the nature of the categories.

One-Hot Encoding

Ideal for **nominal** (unordered) categories, creating new binary features for each category.

Used for multiclass columns.

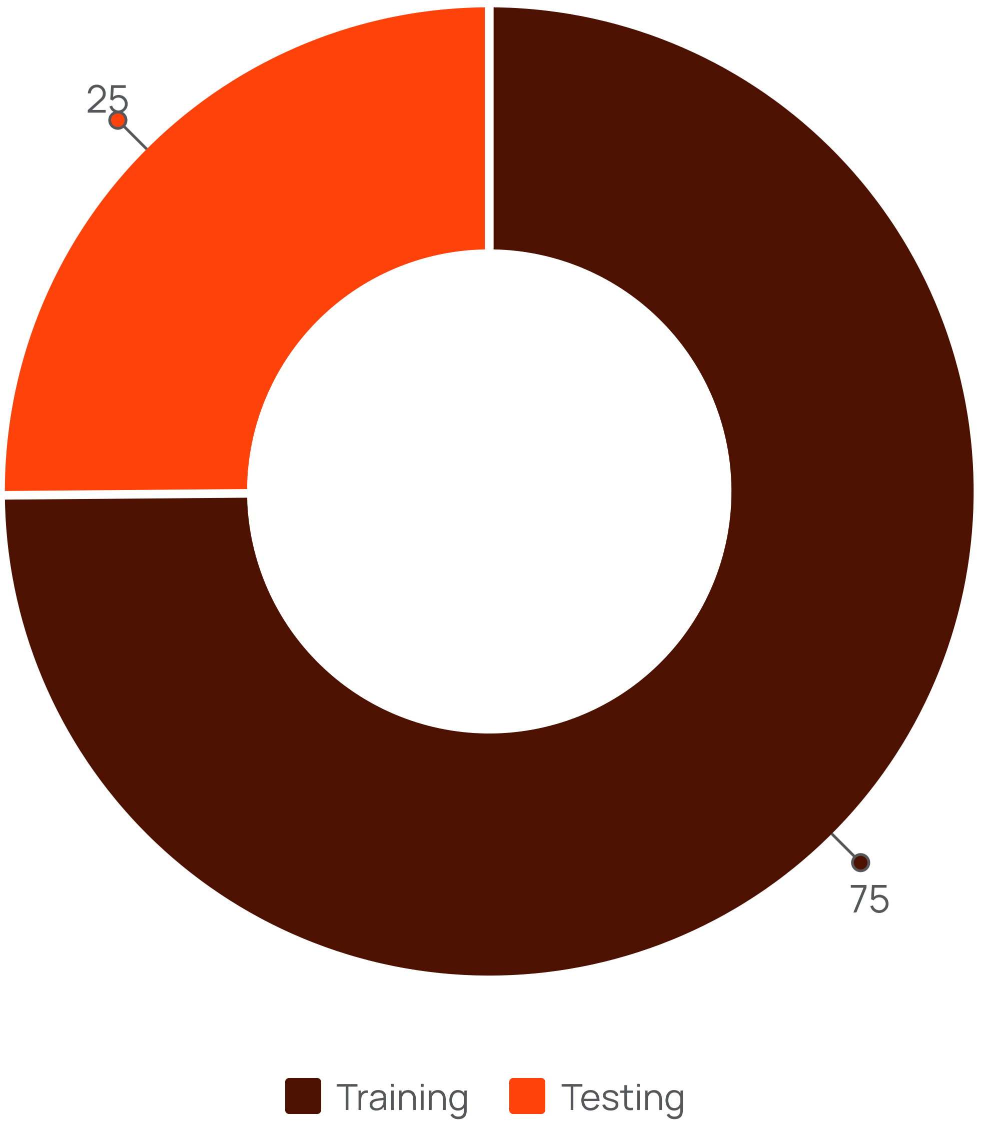
```
pd.get_dummies(newdf[col], prefix=col,
drop_first=True)
```

Label or Ordinal Encoding

Suitable for **ordinal** (ordered) categories, assigning a unique integer to each.

Used for binary class columns

```
label_encoder.fit_transform(newdf[col])
```



Using sklearn.model_selection.train_test_split.

Modeling and Accuracy Calculation for Booking Status

Classification Models

using **Logistic Regression** we get accuracy of 76% after handling imbalanced data.

Evaluation Metrics

Assess performance using: **Accuracy**, **Precision**, **Recall**, **F1-score** in Classification report.

Handling Imbalanced Data

Using **SMOTE** technique to oversample the minor class.

Summary and Best Practices for Data Science Workflows

1

Thorough Data Cleaning

Address nulls, verify data types, and cleanse whitespace.

2

Prudent Outlier Handling

Utilize Z-score methods for detection and proper management.

3

Effective Feature Engineering

Craft meaningful features and correctly encode categorical variables.

4

Data Splitting

Ensure balanced train-test sets for unbiased model evaluation.

5

Robust Model Selection & Evaluation

Choose appropriate binary classifiers and use multiple metrics for assessment.