**Hospital Stay Duration Prediction Report**



**1. Objective**

This project aims to:

* **Task 1**: Perform a comprehensive exploratory data analysis (EDA) on the hospital dataset to extract insights.
* **Task 2**: Build a machine learning model to predict a patient's **length of stay** in the hospital using various attributes of admission and hospital infrastructure.

**2. Dataset Description**

* **Total Records**: 318,000
* **Total Features**: 18 (17 predictors, 1 target)
* **Target Variable**: Stay (length of stay as categorical classes)

**Columns**

| **Feature** | **Description** |
| --- | --- |
| case\_id | Unique ID for the case |
| Hospital\_code | Hospital code |
| Hospital\_type\_code | Type of hospital (categorical: a-g) |
| City\_Code\_Hospital | City code of the hospital |
| Hospital\_region\_code | Region (X, Y, Z) |
| Available\_Extra\_Rooms\_in\_Hospital | Number of extra rooms |
| Department | Medical department (gynecology, surgery, etc.) |
| Ward\_Type | Ward type |
| Ward\_Facility\_Code | Facility code for ward |
| Bed\_Grade | Grade of the bed (1–4) |
| patientid | Unique patient ID |
| City\_Code\_Patient | City code of the patient |
| Type\_of\_Admission | Type (Trauma, Emergency, Urgent) |
| Severity\_of\_Illness | Severity of the condition |
| Visitors\_with\_Patient | No. of visitors |
| Age | Age group (binned) |
| Admission\_Deposit | Deposit made on admission |
| Stay | **Target variable** – number of days of stay (11 categories) |

**3. Preprocessing**

**Missing Values Handling**

* Bed\_Grade: 113 nulls → imputed using **mode**.
* City\_Code\_Patient: 4300 nulls → imputed using **KNN Imputer**.

**Encoding**

* Applied **OrdinalEncoder** to all categorical columns, ensuring consistent mapping for training and inference.

**Target Transformation**

Original Stay classes (11):

['0-10', '11-20', '21-30', '31-40', '41-50', '51-60',

'61-70', '71-80', '81-90', '91-100', 'More than 100 Days']

Converted to midpoint buckets:

bucket\_map = {

'0-10': 5,

'11-20': 15,

'21-30': 25,

'31-40': 35,

'41-50': 45,

'51-60': 55,

'61-70': 65,

'71-80': 75,

'81-90': 85,

'91-100': 95,

'More than 100 Days': 101

}

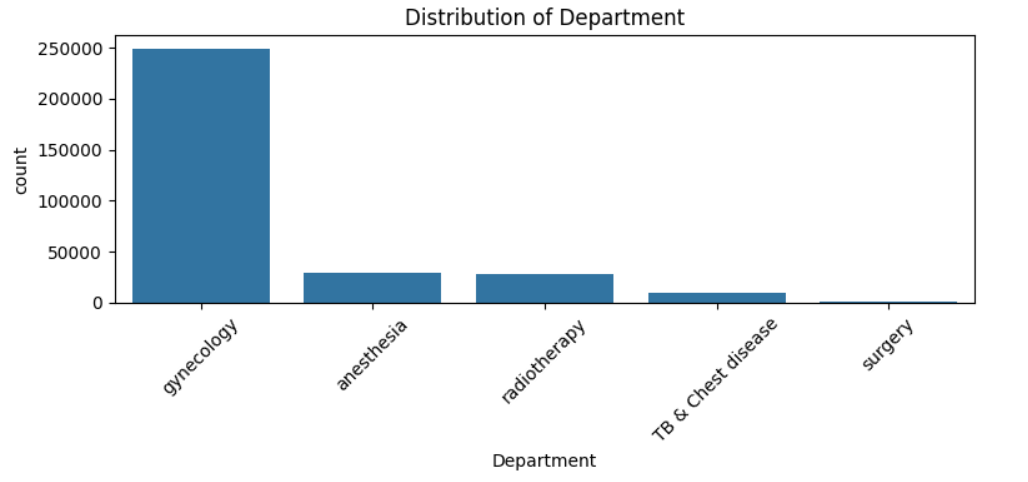
Then re-binned into 5 final categories:

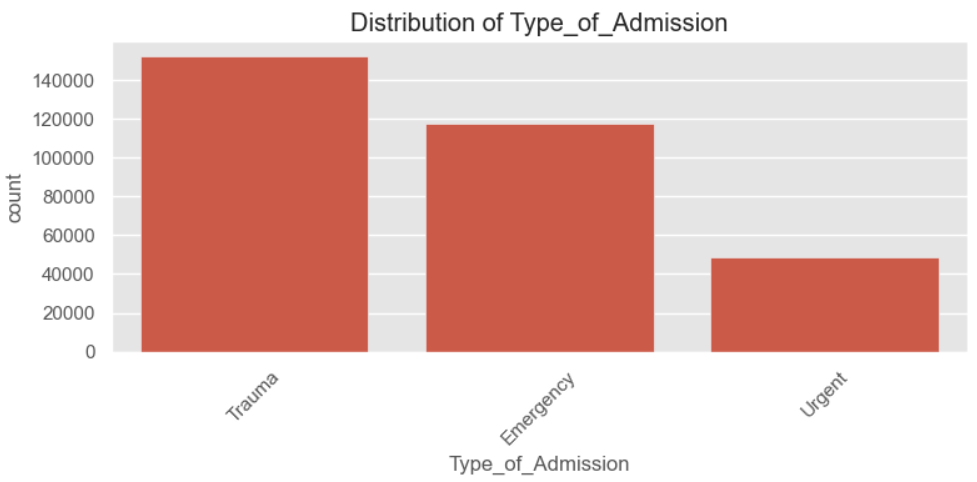
* '0-20'
* '21-40'
* '41-60'
* '61-80'
* '81+'

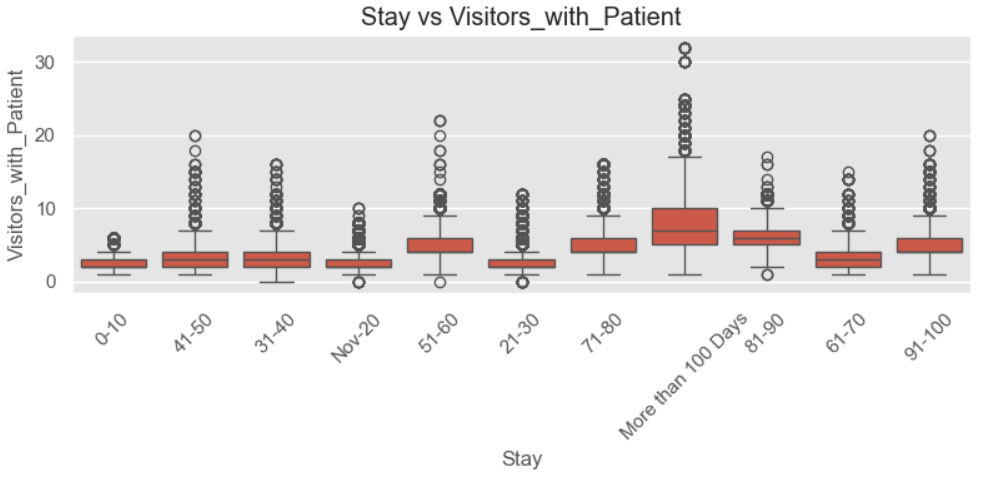
**4. Exploratory Data Analysis (EDA)**

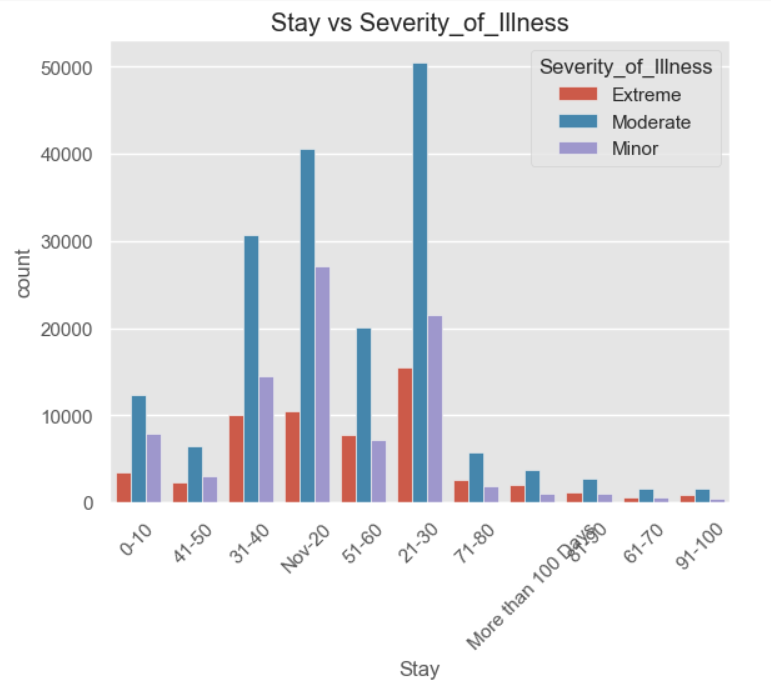
**Distributions**

* **Hospital\_type\_code**: Mostly type a, b, and c
* **Hospital\_region\_code**: Region X dominates
* **Department**: gynecology is by far the most frequent
* **Ward\_Type**: R, Q, and S are dominant
* **Type\_of\_Admission**: Trauma > Emergency > Urgent
* **Severity\_of\_Illness**: Most patients are Moderate
* **Age Groups**: 31–40, 41–50 are the most common



****

****

****

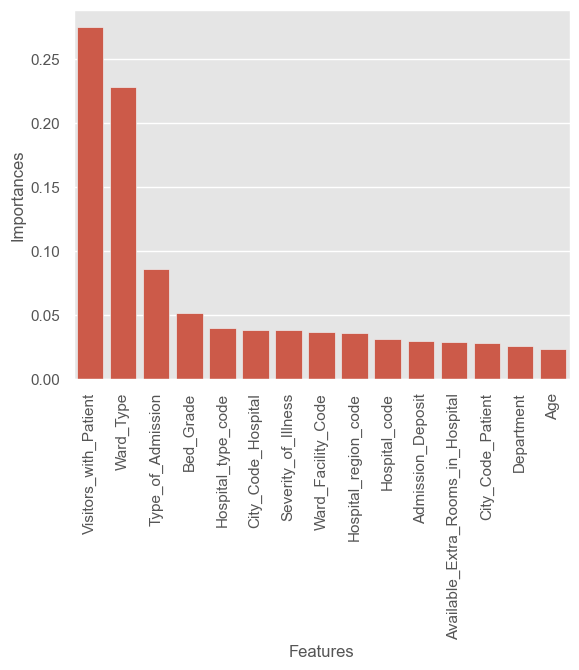
**Target Imbalance**

* Stay has heavy skew:
  + Most common: 21–30 days (27%)
  + Long stays (81+ days): < 3%

**5. Machine Learning Modeling**

**Baseline Model: RandomForestClassifier**

* Input: original 11-class target
* Metrics (Sample):
  + 21–30: F1 = 0.52
  + 31–40: F1 = 0.29
  + More than 100 Days: F1 = 0.48
* Poor performance on underrepresented classes (61–70, 71–80)
* FEATURE IMPORTANCE



**Target Binning Impact**

After grouping into 5 classes:

| **Class** | **Precision** | **Recall** | **F1-score** |
| --- | --- | --- | --- |
| 0-20 | 0.53 | 0.50 | 0.51 |
| 21-40 | 0.55 | 0.65 | 0.59 |
| 41-60 | 0.41 | 0.34 | 0.37 |
| 61-80 | 0.27 | 0.07 | 0.11 |
| 81+ | 0.58 | 0.45 | 0.51 |

**Final Model: XGBoost with Hyperparameter Tuning**

* **Approach**: GridSearchCV with 5-fold CV and 25 parameter combinations
* Improved performance especially for:
  + 21–40: F1 ↑
  + 81+: F1 ↑
  + 0–20: F1 ↑

*XGBoost handled imbalance better and generalized well across all buckets.*

**6. Final Model Selection**

**Model Chosen: XGBoostClassifier**

* + Better performance on skewed target
  + Handles high-cardinality and categorical data well
  + Less overfitting than RF due to regularization
  + Better class-level recall on minority classes (81+, 61–80)

**7. Challenges Faced**

* **Class Imbalance**: Severe skew in original target caused poor recall on minority classes.
* **Overfitting**: RandomForest overfitted without generalization.
* **Data Quality**: Missing values in key features (handled through mode & KNN imputation).
* **High Cardinality**: Some features like patient ID were not useful and dropped.

**8. Conclusion**

* EDA revealed crucial insights into the nature of admissions and hospital operations.
* Binning the target variable significantly improved classification performance.
* XGBoost, with tuned hyperparameters, delivered the best F1-score across classes.
* The model can serve as a clinical decision-support system to estimate patient stay duration based on admission parameters.