

Predicting Patient Length of Stay via Machine Learning and Correlation Analysis: A Data-Driven Approach to Optimize Hospital Resource Allocation and Operational Efficiency

Project implementation - Objective 1 - 4

In this document, we will discuss our Objective 1 - 4, the methods we used to perform our analysis, and the results we achieved, with proper evaluation. Starting from stating our aim, objectives 1 - 4 aimed to achieve the following:

Objective 1

Identification of some external variables that present a positive correlation with patients' length of stay: our first goal is study the correlation between meaningful external variables regarding patient's basic medical information and LOS, extrapolating the relevant ones and giving an explanation to why they present a positive or negative correlation.

Objective 2

Creation of a prediction model: through our first dataset, we aim to create a model that hospitals can use to accurately predict the expected LOS by inputting the relevant patient's medical information.

Objective 3

Identification of all internal variables that present a positive correlation with patients' length of stay: our third goal is study the correlation between all the available variables regarding body levels / mental features and LOS, extrapolating the relevant ones and giving an explanation to why they present a positive or negative correlation.

Objective 4

Creation of a prediction model: through our second dataset, we aim to create a model that hospitals can use to accurately predict the expected LOS by inputting the relevant body levels values.

Objective 5

Recommendations to increase the hospital efficiency: once the models are created and ready for use by the hospitals, we will provide recommendations on how to use such models, and the consequent adjustments that are expected to increase efficiency based on the models result.

The following discussion will cover the methods used during our analysis for objectives 3 and 4. Objective 5 is purely interpretative (no methods or results needed) and will be covered in the final submission.

Before proceeding, as we are using an additional dataset for objectives 3 and 4, below is a descriptive table of our variables.

Variable	Data Type	Description	Mean	Min	Max
lengthofstay	integer	Length of hospital stay (in hours)	4.00103	1	17
malnutrition	binary	Presence of malnutrition (1 = yes, 0 = no)	0.04948	0	1
dialysisrenalendstage	binary	Dialysis or end-stage renal disease (1 = yes, 0 = no)	0.03642	0	1
irondef	binary	Iron deficiency (1 = yes, 0 = no)	0.09494	0	1
pneum	binary	Pneumonia diagnosis (1 = yes, 0 = no)	0.03945	0	1
substancedependence	binary	Substance dependence diagnosis (1 = yes, 0 = no)	0.06306	0	1
psychologicaldisordermajor	binary	Major psychological disorder (1 = yes, 0 = no)	0.23904	0	1
depress	binary	Depression diagnosis (1 = yes, 0 = no)	0.05166	0	1
psychother	binary	Received psychological therapy (1 = yes, 0 = no)	0.04939	0	1
hemo	binary	Hemoglobin level	0.08	0	1
hematocrit	float	Hematocrit percentage	11.97594	4.4	24.1
neutrophils	float	Neutrophil count	10.17746	0.1	245.9
sodium	float	Sodium level (mEq/L)	137.8914	124.91263	151.38728
bloodureanitro	float	Blood urea nitrogen level	14.09718	1.0	682.5
creatinine	float	Creatinine level (mg/dL)	1.099349	0.2197695	2.0352021
pulse	Integer	Pulse rate (beats per minute)	73.44472	21	130
respiration	float	Respiration rate (breaths per minute)	6.493768	0.2	10.0
bmi	float	Body Mass Index (BMI)	29.80576	21.992683	38.935293
glucose	float	Glucose level (mg/dL)	141.9634	1.0059269	271.44428

Methods

Methods for Objective 1

For objective 1, we hope to use our first dataset to explore the relationship between some meaningful independent variables and the length of hospital stay of patients. To achieve this goal, we selected the following variables as independent variables: “Severity of illness”, “Visitors With Patient”, “Admission Deposit”, “Age”. The dependent variable is “Stay”.

After selecting the independent and dependent variables, we first conducted exploratory data analysis. First, analyze whether each variable has missing values and choose an appropriate method to fill them. Next, we encode the ordered categorical variable "Severity of illness" with 1-3 corresponding to the severity of the disease. At the same time, since the "Age" and "Stay" variables are binned, we randomly select a value from the corresponding level for the data to be used in the subsequent regression. After processing the data, we draw frequency distribution histograms for "Admission Deposit", "Visitors with Patient", and the corresponding variables "Age" and "Stay" generated after random sampling and draw a bar chart for the variable "Severity of Illness". Through this method, we can understand the basic situation of the data.

After completing EDA, we performed linear regression, assuming that the independent variables and dependent variables conform to the following model:

$$LOS = \beta_0 + \beta_1 \cdot \text{Illness_code} + \beta_2 \cdot \text{Visitors_with_Patient} + \beta_3 \cdot \text{Admission_Deposit} + \beta_4 \cdot \text{Age_Value}$$

We use OLS linear regression to fit a straight line, obtain the influence coefficient and intercept of different independent variables, and analyze the results.

Methods for Objective 2

For objective 2, our goal is to use the first dataset to build a model that can be used by hospitals to predict the length of patient stay, thereby improving the efficiency of hospital management.

Because the data in the first dataset is a combination of continuous data and categorical data, and the final prediction variable is a binned variable, that is, label. So, we naturally choose the classification model in machine learning to make predictions. To use machine learning to build a model, we first determine X and y. X is all variables except "Stay", and y is "Stay". After that, we use 80% of the data as a training set and 20% of the data as a test set for subsequent model building and evaluation.

After dividing the training set and the test set, we perform exploratory data analysis on the training set data. We draw distribution graphs for "Admission Deposit", "Visitors with patient", and draw histograms for "Age", "Stay", "Type of Admission", and "Severity of Illness". Finally, we encode the categorical variables, calculate the correlation matrix together with the continuous variables, and draw a heatmap to show the correlation between different variables.

After completing this step, we start feature engineering. First, we adjust the variables with inappropriate data types to the most appropriate data types, such as converting "Age" to string type. Next, discard the noise variables ('case_id', 'patientid', 'City_Code_Hospital', 'Ward_Facility_Code', 'City_Code_Patient'), and keep the variables that are useful for building the model. Next, for continuous variables, we use the median to fill missing values and perform standardization; for categorical variables, we use the mode to fill missing values and perform coding for subsequent model construction. After completing this, we checked the classification of different categories of y to see if the samples became more balanced after undersampling. We also used SMOTE (Synthetic Minority Over-sampling Technique) to perform oversampling to ensure that the number of samples for all categories becomes the same, thus solving the class imbalance problem.

After feature engineering, we can start formal model building. In this study, we first use logistic regression to build the model. At the same time, we use the 5-fold cross-validation method to divide the training set into 5 parts, 4 for training and 1 for validation, and repeat it 5 times. We record the accuracy of each cycle and calculate the average accuracy.

After training the model, we enter the model evaluation stage. First, we calculate the basic classification indicators (Accuracy, Precision, Recall, F1), calculate the multi-classification

probability output, and use `roc_auc_score()` to calculate the OVO multi-class AUC. Then use `confusion_matrix()` and `seaborn.heatmap()` to draw the confusion matrix to show the misclassification of each category prediction. Perform One-vs-Rest PR curve analysis on each category to observe the precision performance of each category at different recall levels. Finally, use `label_binarize()` to convert the multi-class labels into OvR format, draw the ROC curve of each category and calculate the AUC. The AUC score of each category is shown in the legend to directly quantify the "discrimination ability" of each category.

Methods for Objective 3

For Objective 3, our primary goal was to estimate the correlation between the internal variables in our dataset and the dependent variable, Length of Stay. Understanding these relationships is critical to identifying which variables may be useful predictors in our regression model and which may have limited explanatory power.

To perform this analysis, we used the R programming language in conjunction with the Radiant interface, a visual analytics tool that streamlines statistical tasks. We began by installing the Radiant package in R, ensuring that all required dependencies were properly set up. Once installed, we launched the Radiant software, which provides a web-based interface for performing statistical operations on datasets.

Within Radiant, we navigated to the Data Management window, where we uploaded our complete dataset in CSV format. The dataset included 100,000 observations, representing a rich and comprehensive set of clinical records. After successfully uploading the file, we transitioned to the Data > Basics section of the software.

Here, we configured the analysis by setting the table type to “Correlation”, which enabled us to compute pairwise relationships between variables. Since our dataset consists predominantly of numerical variables, we selected the Pearson correlation method, which is appropriate for measuring the strength and direction of linear relationships between continuous variables.

Next, we selected a group of X-variables of particular interest and calculated their correlation coefficients with Length of Stay. Radiant’s visualization tools made it easy to view the correlations in both tabular and graphical format, providing an immediate view of which predictors might have meaningful linear associations with the outcome variable.

As said, our primary focus was to examine the relationship between each internal variable and the target variable, Length of Stay (LOS). This analysis aimed to identify which physiological and psychological features in the dataset show a positive or negative association with patients’ hospitalization time.

To structure the analysis, we classified the independent variables into two main categories:

- Unordered variables, which include nominal categorical variables such as diagnosis indicators or binary conditions (e.g., presence of depression, substance dependence)

- Ordered variables, which include continuous numeric variables (e.g., sodium level, glucose, BMI) and ordinal categorical variables if applicable.

For unordered variables, we used the correlation ratio to estimate the strength of association with LOS. An initial p-value for this type of correlation isn't provided, as the correlation ratio is an effect size measure derived from ANOVA (Analysis of Variance), not a statistical test by itself. It tells us how much variance in the dependent variable is explained by group membership but not whether that difference is statistically significant. In order to associate p-values to each correlation ratio, we performed a one-way ANOVA test for each categorical variable against the outcome variable (Length of Stay).

For ordered variables, we calculated the Spearman rank correlation coefficient, which is more robust for detecting monotonic (not necessarily linear) relationships and less sensitive to outliers than Pearson correlation.

This dual approach allowed us to accurately quantify the extent and direction of association between a wide range of internal patient features and LOS.

The results of this correlation analysis will be presented and interpreted in detail in the *Results* section of this report.

Methods for Objective 4

Objective 4 required significant more planning and thinking. As our goal is to create a prediction model, the formulation of a regression formula was necessary. For this purpose, we analyzed the results of the correlation analysis, and then proceeded to include all the variables that showed a low enough p-value to be significant as a first step. The initial regression formula came to be:

$$\begin{aligned} \text{Length of Stay} = & \beta_0 + \beta_1 \cdot \text{malnutrition} + \beta_2 \cdot \text{dialysisrenalendstage} + \beta_3 \cdot \text{irondef} + \beta_4 \cdot \text{pneum} \\ & + \beta_5 \cdot \text{substancedependence} + \beta_6 \cdot \text{psychologicaldisordermajor} + \beta_7 \cdot \text{depress} + \beta_8 \\ & \cdot \text{psychoter} + \beta_9 \cdot \text{hemo} + \beta_{10} \cdot \text{bloodureanitro} \end{aligned}$$

After analysing this formula, we realized that a lot of supposedly significant variable were left out. In order to address this problem, we proceeded with studying each one of the variables and assess if they could have become part of the regression model through transformations. Initially, we focused on sodium levels and bmi. For sodium, we realized that both very low or very high sodium levels are associated with serious health issues like hyponatremia or hypernatremia, both of which can prolong hospitalization. This suggested the presence of a U-shaped relationship, bringing us to creating a quadratic term that captures the idea that both too little and too much sodium are bad. Similarly, the quadratic term for bmi indicates that both underweight and obesity may lead to longer stays, again forming a nonlinear (U-shaped) pattern.

Coming to the possible interaction variables, we reasoned that depression and major psychological disorders usually interact to worsen outcomes. Patients with both likely face more

complex care coordination, slower recovery, and medication management challenges. The interaction variables is explained by the combination having a stronger impact than either condition alone. Similarly, the interaction hemo x bloodureanitrogen captures the interaction between anemia (hemo) and kidney function (BUN), since if someone has both low hemoglobin and poor kidney function, it's a sign of systemic illness, which likely extends length of stay.

Including these variables, we came to an updated regression model:

$$\begin{aligned} \text{Length of Stay} = & \beta_0 + \beta_1 \cdot \text{malnutrition} + \beta_2 \cdot \text{dialysisrenalendstage} + \beta_3 \cdot \text{irondef} + \beta_4 \cdot \text{pneum} \\ & + \beta_5 \cdot \text{substance dependence} + \beta_6 \cdot \text{psychological disordermajor} + \beta_7 \cdot \text{depress} + \beta_8 \\ & \cdot \text{psychoter} + \beta_9 \cdot \text{hemo} + \beta_{10} \cdot \text{bloodureanitro} + \beta_{11} \cdot \text{bmi} + \beta_{12} \cdot \text{bmi}^2 + \beta_{13} \\ & \cdot \text{sodium} + \beta_{14} \cdot \text{sodium}^2 + \beta_{15} \cdot (\text{depress} \times \text{psychological disordermajor}) + \beta_{16} \\ & \cdot (\text{hemo} \times \text{bloodureanitro}) \end{aligned}$$

However, we realized that the transformations we just performed, in the medical field, are often required, as body levels and conditions often present in different scales, magnitudes, and interacting with each other. For these reasons, we continued our analysis of the remaining variables to ensure nothing significant was being left out. Starting from hematocrit, the inverse has been taken to capture its nonlinear, inverse relationship: as hematocrit decreases, the risk of complications and longer hospital stay rises sharply, but only when it's low enough. The inverse transformation exaggerates the effects at low hematocrit levels, where clinical impact is greater. About neutrophils, neutrophil counts are usually right-skewed, as a few patients can have very high levels due to infection or inflammation. The rationale behind including this variable is that elevated neutrophils are markers of acute infection or inflammation, which can increase hospital stay. Taking the log dampens the influence of extreme values and better models diminishing returns: a small increase at low levels may matter more than a large one at already-high levels. Moving to glucose, taking the squared term derived from the presence of, again, a u-shaped relationship, as both hypoglycemia (low blood sugar) and hyperglycemia (high blood sugar) can lead to complications. For creatinine a logarithmic function has been used. Like neutrophils, creatinine is right-skewed and benefits from log transformation to model its relationship more proportionally. It's important to include it as creatinine is a key indicator of kidney function. Impaired kidneys can complicate treatment, medication dosing, and discharge planning. For pulse, the relationship between pulse and length of stay may be nonlinear: risk increases sharply at low pulse rates, but levels off at higher rates, reason why the. Inverse function has been applied. Abnormal pulse is a proxy for cardiovascular or metabolic instability, especially early in a hospital stay. The respiration has been squared, to capture the idea that both abnormally low and high respiration rates can be concerning. Its interaction variable with pneumonia is relevant as patients with pneumonia may only show a problematic increase in length of stay if they also have abnormal respiration. This expresses

how pneumonia severity (captured by breathing rate) influences outcome, as it depends on the respiratory response.

After all these considerations and transformations, our final updated model presents as:

$$\begin{aligned} \text{Length of Stay} = & \beta_0 + \beta_1 \cdot \text{malnutrition} + \beta_2 \cdot \text{dialysisrenalendstage} + \beta_3 \cdot \text{irondef} + \beta_4 \cdot \text{pneum} \\ & + \beta_5 \cdot \text{substancedependence} + \beta_6 \cdot \text{psychologicaldisordermajor} + \beta_7 \cdot \text{depress} + \beta_8 \\ & \cdot \text{psychoter} + \beta_9 \cdot \text{hemo} + \beta_{10} \cdot \text{bloodureanitro} + \beta_{11} \cdot \text{bmi} + \beta_{12} \cdot \text{bmi}^2 + \beta_{13} \\ & \cdot \text{sodium} + \beta_{14} \cdot \text{sodium}^2 + \beta_{15} \cdot (\text{depress} \times \text{psychologicaldisordermajor}) + \beta_{16} \\ & \cdot (\text{hemo} \times \text{bloodureanitro}) + \beta_{17} \cdot \text{hematocrit} + \beta_{18} \cdot \text{neutrophils} + \beta_{19} \\ & \cdot \text{creatinine} + \beta_{20} \cdot \text{pulse} + \beta_{21} \cdot \text{respiration} + \beta_{22} \cdot \frac{1}{\text{hematocrit}} + \beta_{23} \\ & \cdot \log \text{neutrophils} + \beta_{24} \cdot \text{glucose} + \beta_{25} \cdot \text{glucose}^2 + \beta_{26} \cdot \log \text{creatinine} + \beta_{27} \\ & \cdot \frac{1}{\text{pulse}} + \beta_{28} \cdot \text{respiration}^2 + \beta_{29} \cdot (\text{respiration} \times \text{pneumonia}) \end{aligned}$$

Note how, despite the addition of the transformed variables, we keep all the original ones. This is because the transformed variable alone doesn't capture the full picture. The original variable often still carries unique linear information that the transformation can't fully replace, while the transformed version (e.g., squared, log, inverse) captures the nonlinear patterns. For instance, as sodium has a U-shaped effect on Length of Stay (i.e., both very low and very high values are bad), then: sodium captures the tilt or slope of the curve, while sodium² captures the curvature. If we only include sodium², we miss half of the effect. Similarly, if we only include bmi², the model assumes the relationship must pass through zero when bmi = 0, which doesn't make sense. It becomes less interpretable and may misfit the data. This is because if we included just the transformed variable, the model assumes the relationship with the outcome has no linear component, which is often not true. This can lead to incorrect estimates, poor fit, and misleading conclusions.

Software implementation

In Objective 1, we used Python.

In Objective 2, we used Python.

In Objective 3, we used Python, R, and Radiant.

For Objective 4, we used R and Radiant.

Results

Objective 1

For this objective, we need to use OLS regression to find out the impact of the variables we selected on length of stay. By using Python analysis, we can get the following results:

OLS Regression Results						
=====						
Dep. Variable:	LOS	R-squared:	0.292			
Model:	OLS	Adj. R-squared:	0.292			
Method:	Least Squares	F-statistic:	3.277e+04			
Date:	Thu, 03 Apr 2025	Prob (F-statistic):	0.00			
Time:	12:26:53	Log-Likelihood:	-1.3718e+06			
No. Observations:	318438	AIC:	2.744e+06			
Df Residuals:	318433	BIC:	2.744e+06			
Df Model:	4					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]
const	0.0044	0.203	0.022	0.983	-0.393	0.402
Illness_code	3.1568	0.049	64.982	0.000	3.062	3.252
Visitors with Patient	6.3610	0.018	347.085	0.000	6.325	6.397
Admission_Deposit	0.0007	2.97e-05	22.675	0.000	0.001	0.001
Age Value	0.0452	0.002	26.840	0.000	0.042	0.048
=====						
Omnibus:	43697.392	Durbin-Watson:	1.634			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	77709.760			
Skew:	0.904	Prob(JB):	0.00			
Kurtosis:	4.609	Cond. No.	3.21e+04			
=====						

Notes:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 3.21e+04. This might indicate that there are strong multicollinearity or other numerical problems.

Based on this result, we can make the following analysis:

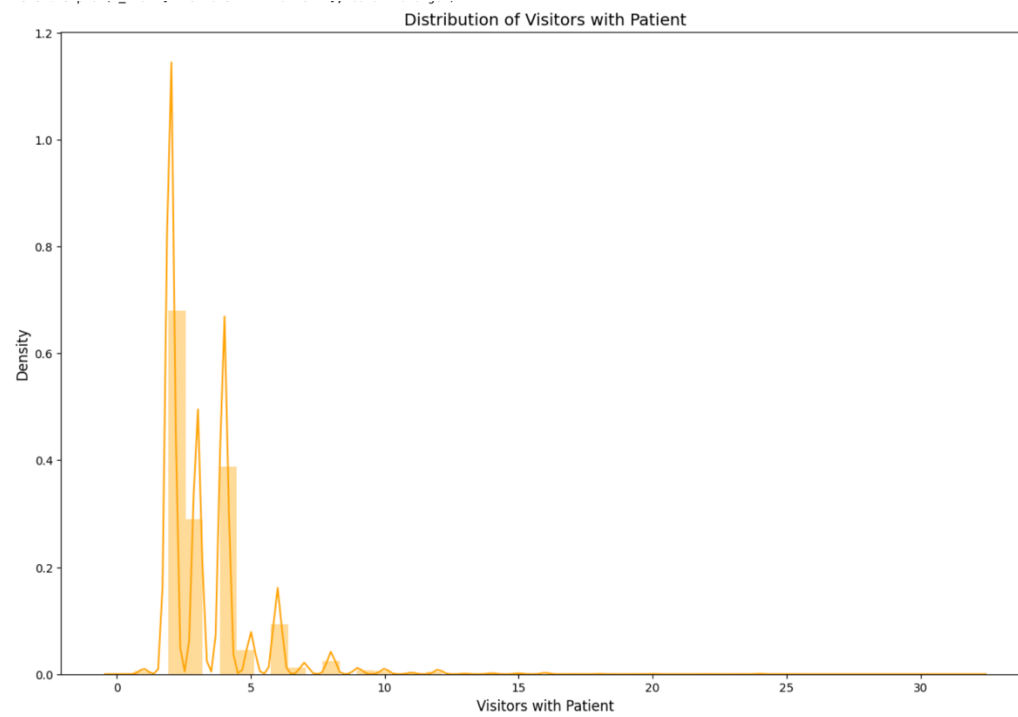
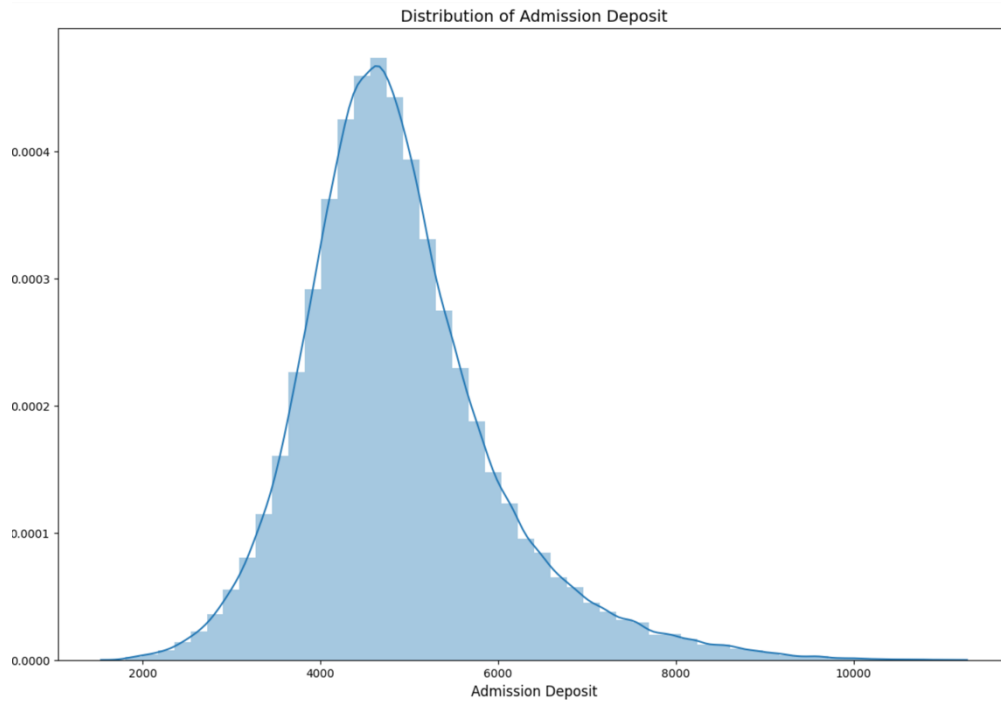
1. The coefficient before illness_code is 3.1568, and the p-value is 0.000, which is statistically significant. This indicates that for every increase in the severity of the illness (such as from moderate to extremely severe), the LOS increases by an average of about 3.16 days, which is very significant.
2. The coefficient before Visitors with Patient is 6.3610, and the p value is 0.000, which is statistically significant, indicating that for each additional visitor, the LOS increases by an average of about 6.36 days, which is a huge and highly significant impact. Further investigation of the underlying mechanism may be needed (is it that more care from relatives leads to a longer treatment cycle? Or is it because of serious illness that there are more visitors?).

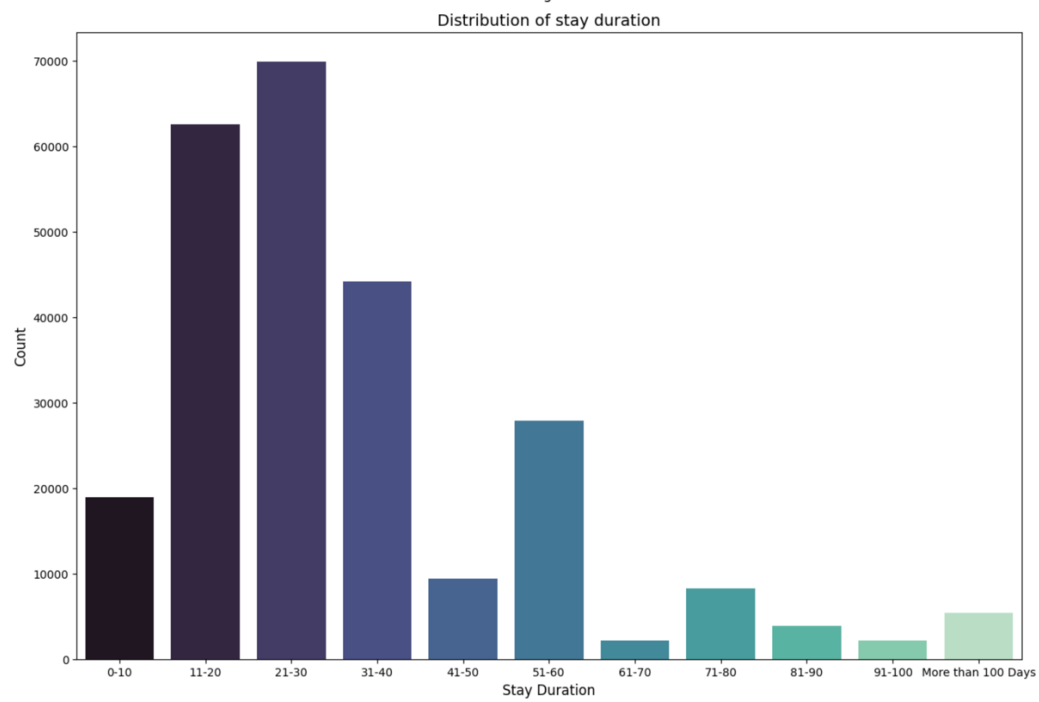
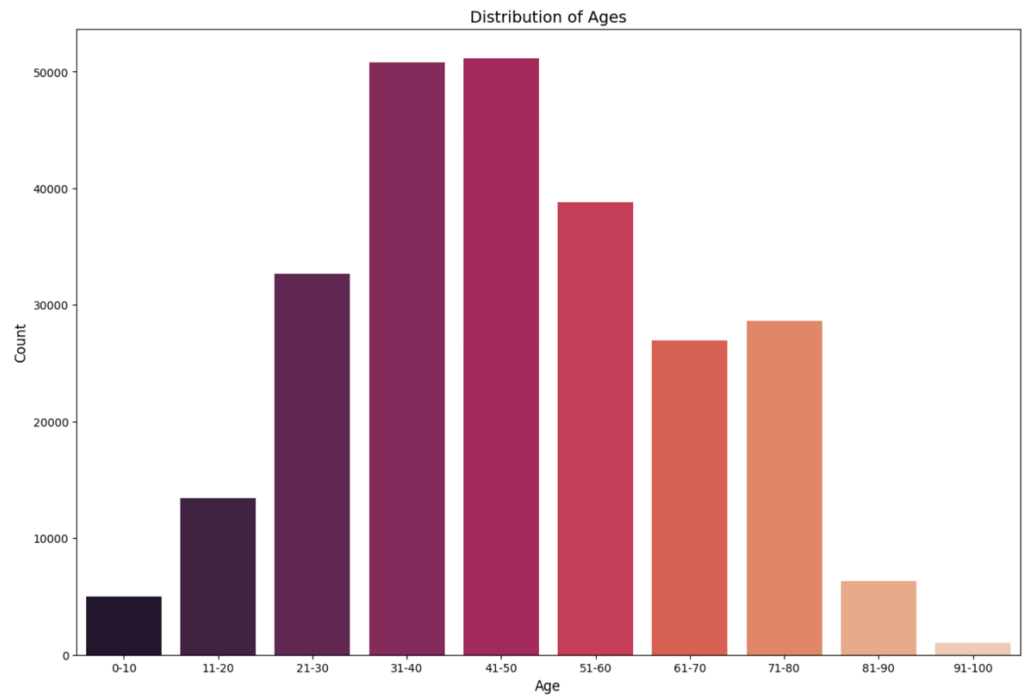
3. The coefficient of Admission Deposit is 0.0007. Although it is very small, the P value is significant, indicating that paying more deposits is related to longer hospital stays, which may be a proxy for medical condition or medical insurance information.
4. The coefficient before Age Value is 0.0452, and the p-value is 0.000, which is statistically significant, indicating that for every additional year of age, LOS increases by 0.045 days, and the elderly have longer hospital stays, which is reasonable in the medical field.
5. The Intercept (const) is 0.0044, and the p-value is 0.983, which is not significant, indicating that the baseline value of LOS is almost meaningless when all variables are 0, which is a normal phenomenon.

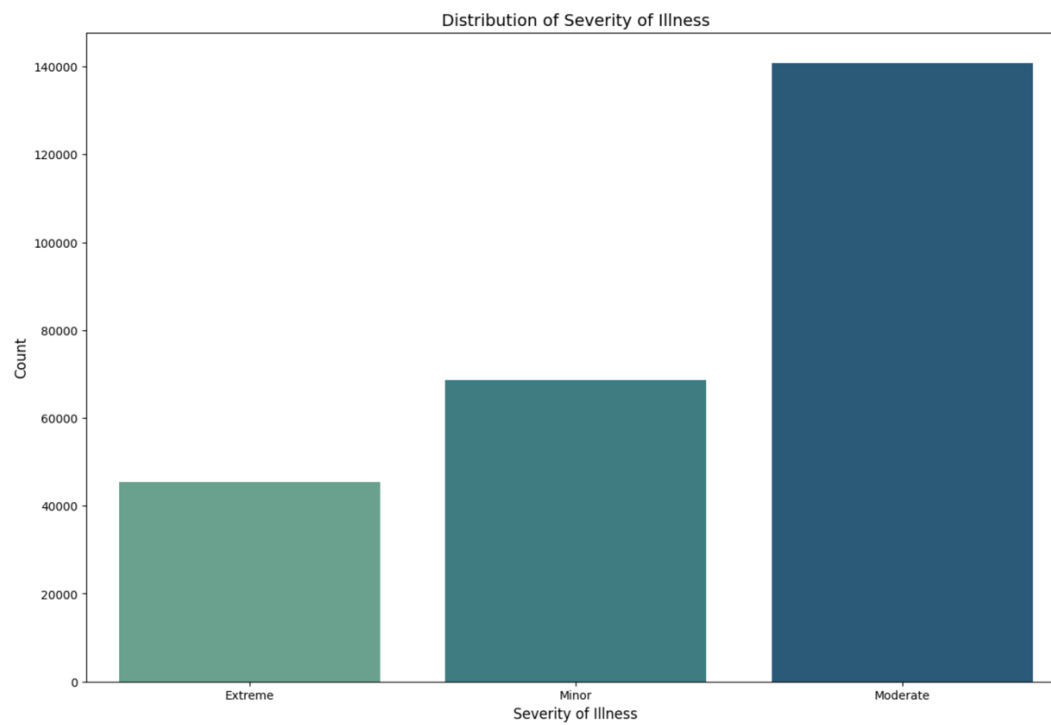
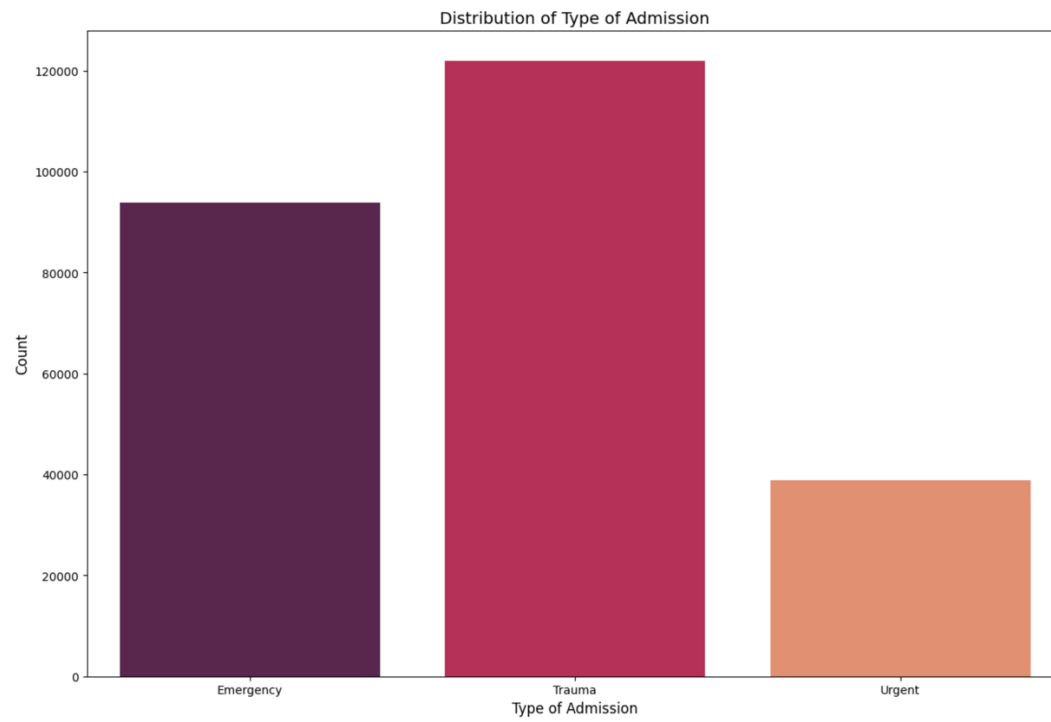
In addition to the above valuable findings, we can see that the model goodness of fit is generally weak, with R squared only 0.292, but it is within an acceptable range; at the same time, the residuals of the model are non-normal, and there may be multicollinearity problems, which can be further studied.

Objective 2

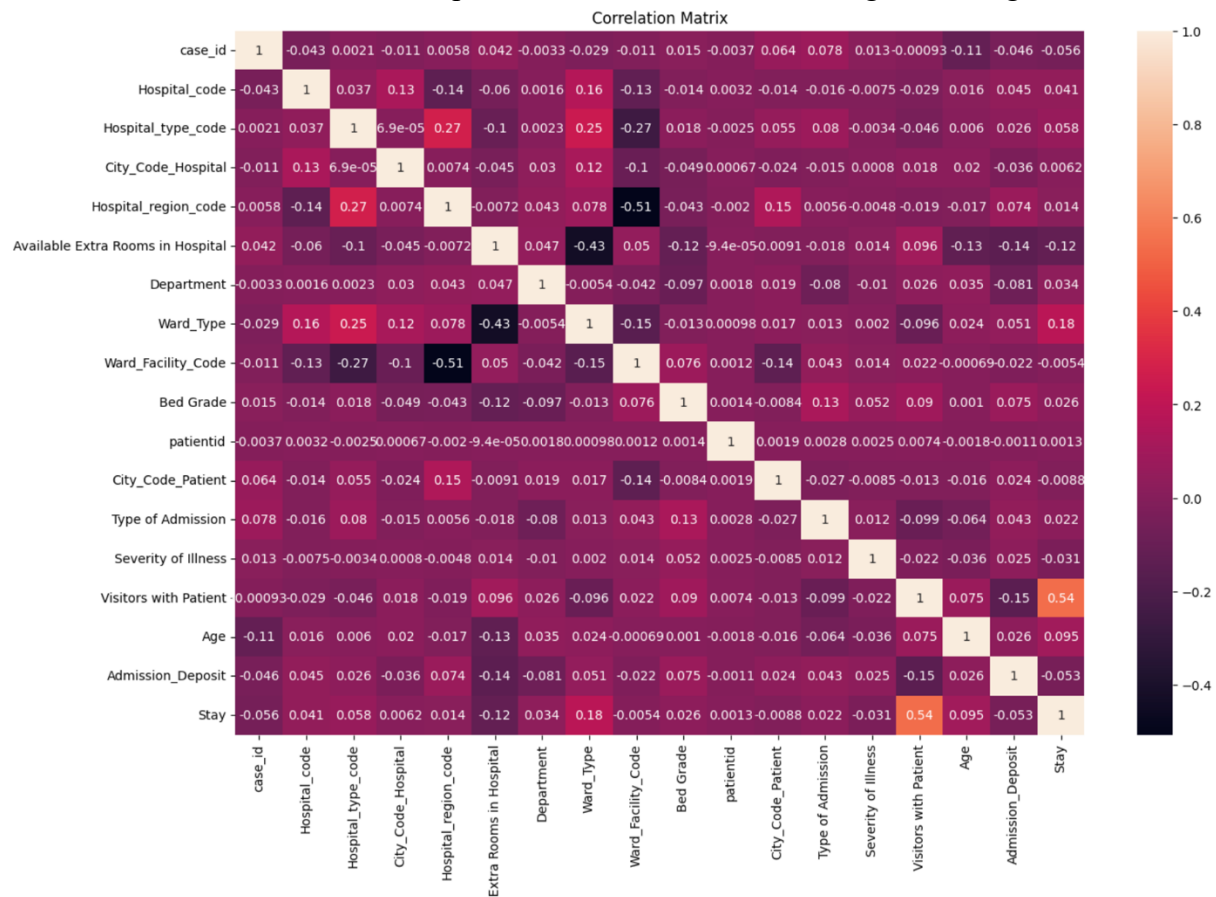
For this Objective, our goal is to build a model that can predict the length of hospital stay of patients based on their medical information. In the initial stage, we conducted an exploratory data analysis and drew the following figure. From the following figure, we can see that Admission Deposit, Visitor with Patient, Age, and Stay are all close to normal distribution, which is consistent with the actual situation.







In addition, we also draw a heatmap of the correlation matrix through encoding.



We can see that there are different correlations between "Stay" and different variables, which will help us with further analysis.

We then performed feature engineering, filled in missing values, encoded categorical variables, and oversampled the samples to ensure that the number of samples in all categories became the same, thus solving the class imbalance problem.

```
Class Distribution After Undersampling:
Stay
21-30      9.090909
More than 100 Days  9.090909
11-20      9.090909
61-70      9.090909
51-60      9.090909
31-40      9.090909
0-10       9.090909
71-80      9.090909
41-50      9.090909
91-100     9.090909
81-90      9.090909
Name: proportion, dtype: float64
```

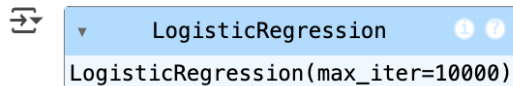
After completing feature engineering, we use the logistic regression method to train the model with the data from the training set.

Model Evaluation

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import cross_val_score

# Logistic regression for classification
model = LogisticRegression(max_iter=10000)

# Fit the model on the training data
model.fit(x_train, y_train)
```



```
LogisticRegression(max_iter=10000)
```

After successful training, we use the test set to evaluate the model. First, we calculate the basic classification indicators (Accuracy, Precision, Recall, F1), calculate the multi-class probability output and use `roc_auc_score()` to calculate the OVO multi-class AUC, and get the following results:

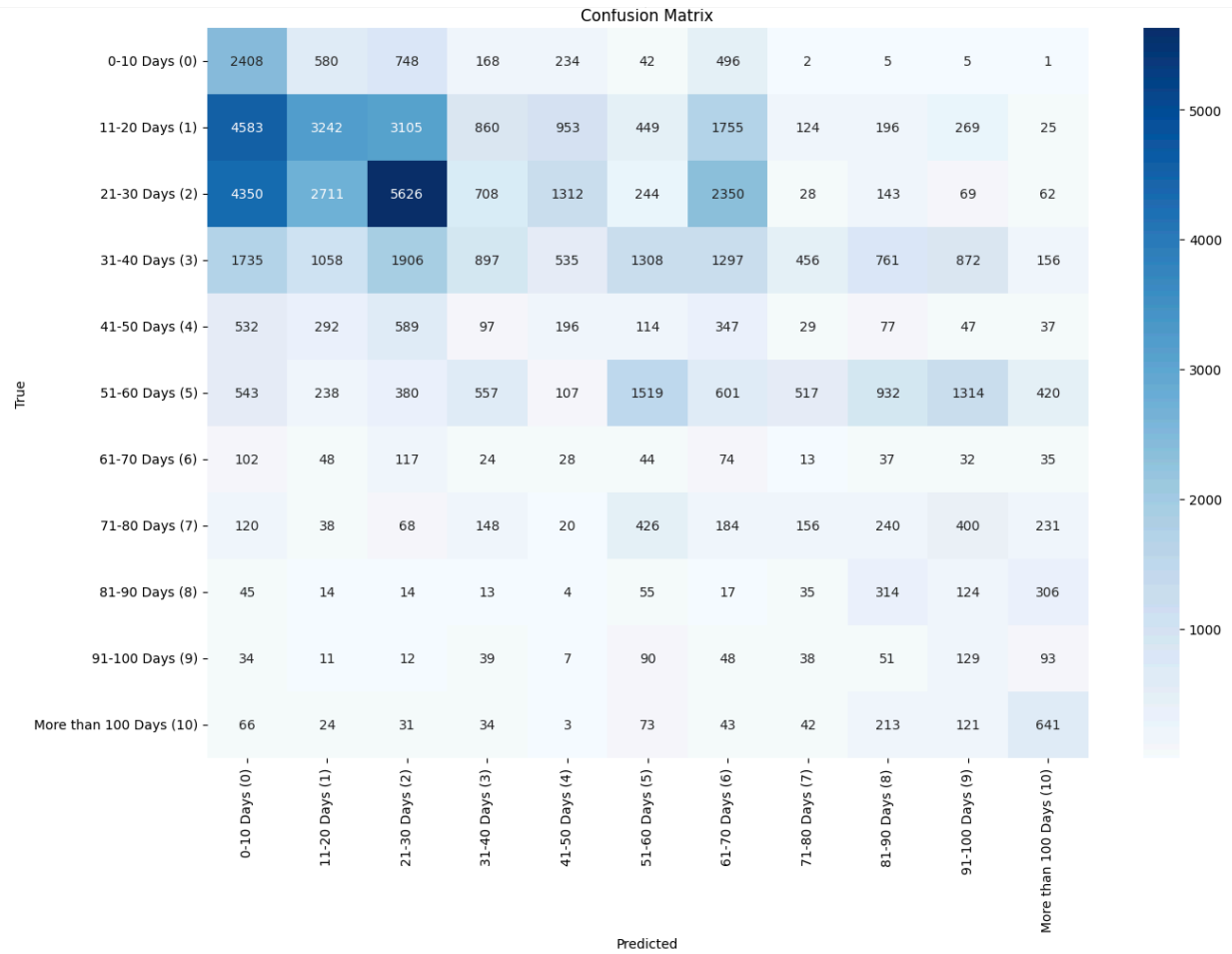
Accuracy: 0.239
Precision: 0.204
Recall: 0.245
F1-score: 0.189
AUC-ROC: 0.734

The following conclusions can be drawn from this result:

1. **Accuracy: 0.239:** The prediction accuracy is only 23.9%, which is slightly higher than random guessing in the 11-category task (random guessing is about $1/11 \approx 9.1\%$). This shows that the model has indeed learned some rules, but the overall prediction ability is still weak.
2. **Precision: 0.204:** Among the samples predicted to be of a certain class, only 20.4% are actually of that class. This means that the model prediction results contain "many false positives" - mistaking other classes for a certain class.
3. **Recall: 0.245:** Only 24.5% of the samples that actually belong to a certain class are correctly identified. This means that the model has a lot of "false negatives", especially in small or easily confused classes.
4. **F1-score: 0.189:** It is a weighted comprehensive indicator of Precision and Recall. The value is very low, indicating that the overall balance is poor. The model is neither accurate nor comprehensive.
5. **AUC-ROC: 0.734:** Although both Accuracy and F1 are low, the AUC reaches 0.734, which is actually a good result. This shows that the model's predicted probability

distribution has strong discriminative ability (although the final classification output is not ideal).

In this case, we have plotted the confusion matrix of the model to further distinguish the causes of errors.

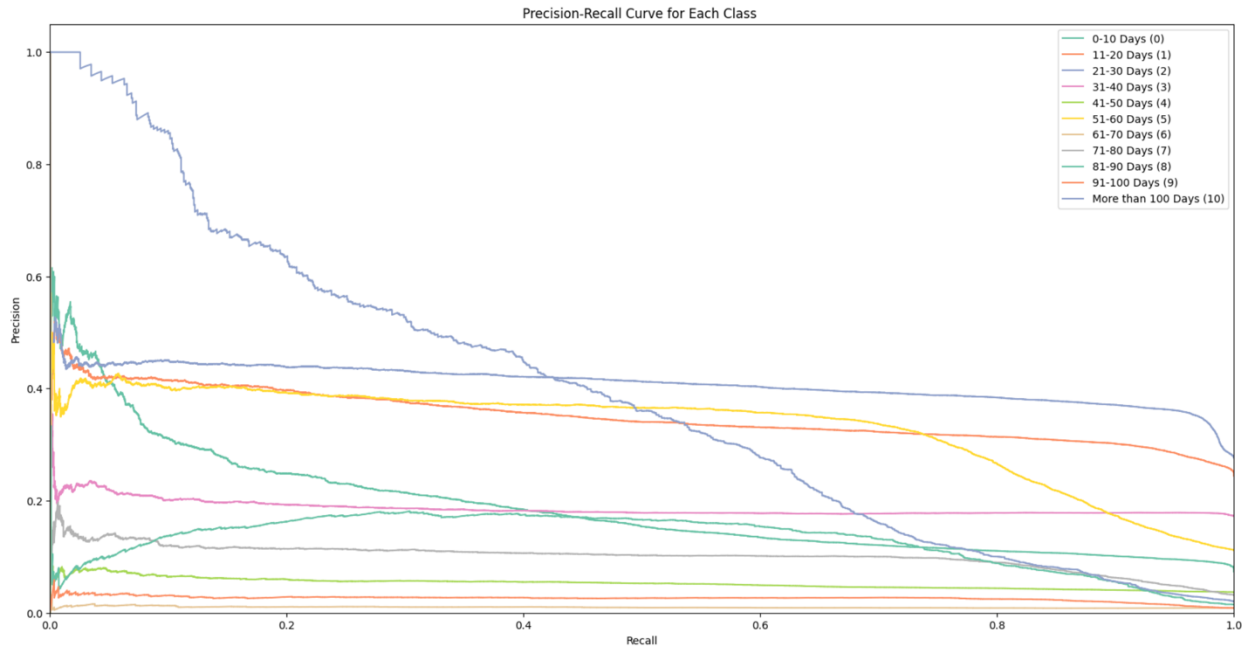


From the confusion matrix we can conclude:

1. **Models with better performance:** Categories 0-10 days, 11-20 days, and 21-30 days, because these categories have many predictions and a high accuracy rate, indicating that the model is good at identifying "short-term hospitalization" situations
2. **Models with poor performance:** Classes 31-70 days (classes 3-6) These classes are almost all confused with other neighboring classes, indicating that these classes are inseparable in feature space, or the features are not distinguishable enough
3. The model has a moderate preference for "long-term hospitalization" (classes 8, 9, and 10). For example, 641 cases of "More than 100 Days" (class 10) were correctly predicted, and most of the misclassifications were concentrated in 8 and 9. This shows that although

the model cannot accurately classify, it can roughly judge that "this type of patient stays longer."

Next, we analyze the Precision-Recall curve of the model and get the following results:



In this image, the horizontal axis (Recall):

$$Recall = TP / (TP + FN)$$

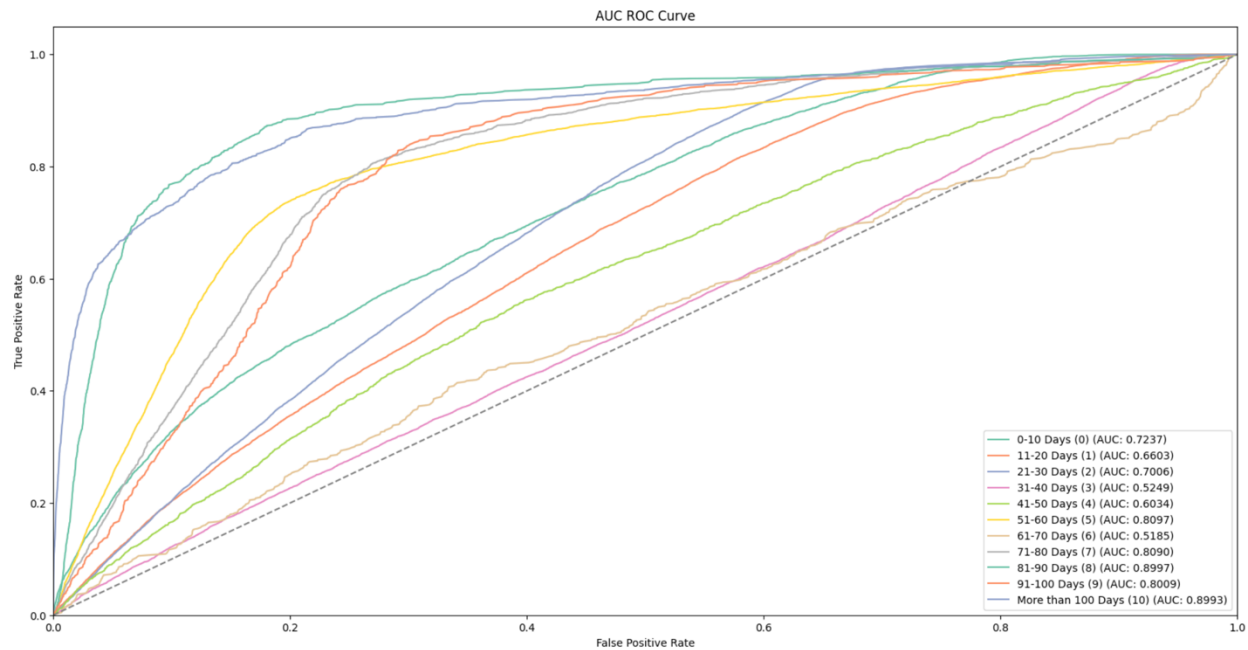
the more to the right, the more positive classes are recognized. Vertical axis (Precision):

$$Precision = TP / (TP + FP)$$

the higher, the fewer false positives. Each curve is the PR curve of the class under the One-vs-Rest (two-classification) setting; the closer the curve is to the upper right corner, the better the prediction performance of the class.

The precision-recall curves show that the classification performance is highly imbalanced across categories. While the model achieves high precision and recall for long-stay categories such as "more than 100 days" and "81-90 days", it performs poorly in medium-stay categories such as "31-40 days" and "61-70 days", where both precision and recall remain very low. This highlights the limited ability of the model to distinguish overlapping patterns of medium-stay durations and suggests the need for feature enhancement, alternative category groupings, or a hierarchical classification strategy.

Finally, we perform multi-classification ROC curve analysis to make more accurate judgments on the predictions of different categories:



This figure is AUC ROC Curve for Each Class, which is used for One-vs-Rest (OvR) ROC curve analysis of each class in a multi-classification model. The quality of each curve reflects the model's ability to distinguish the probability of the class, that is, whether it can give a higher prediction probability to "samples belonging to this class" and a lower probability to "samples not belonging to this class".

In this figure, the horizontal axis (False Positive Rate) represents the false alarm rate, the smaller the better; the vertical axis (True Positive Rate) represents the recall rate, the larger the better; the ideal model is that the closer the curve is to the upper left corner, the better; the gray dotted line is the random guessing baseline (AUC=0.5); the AUC value corresponding to each class is marked in the legend.

Through the curve and AUC value, we can draw the following conclusions:

1. **The strongest prediction categories:** 81-90 days, >100 days, 91-100 days. The model can clearly identify these long-term hospitalization patients and give a high confidence prediction probability.
2. **Fair performance categories:** 51-60 days, 71-80 days, 0-10 days. The model can distinguish these categories from other categories to a certain extent, but the accuracy may not be stable enough
3. **The worst performing categories (special attention is recommended):** 31-40 days (category 3), 61-70 days (category 6). The AUC is close to random guessing (0.5),

indicating that the model has no judgment on the probability output of these two categories. It is strongly recommended to check the sample features and sample size of these two categories, and whether they can be merged, or new features can be introduced.

In summary, the model can make accurate predictions for some categories, but not for others, which can provide a certain degree of reference for hospitals. However, due to the poor performance of the overall prediction index of the model, RandomForestClassifier or XGBoostClassifier can be used to train a new and more accurate model.

Objective 3

This objective focused on identifying internal patient characteristics, both physiological and psychological, that are associated with longer or shorter hospital stays. These variables include body measurements and mental health indicators that may influence patient recovery time and care complexity.

The results revealed a number of variables with notable correlations to length of stay. Some factors, such as low hemoglobin levels or the presence of psychological disorders, showed a clear positive association with longer hospitalization. Others appeared to have more complex or nonlinear relationships. The following section details which variables demonstrated meaningful correlations with LOS, and explores the possible clinical relevance behind the patterns of the apparently non-meaningful ones.

Our Findings

Starting from the correlation table of the ordered variables through Spearman method (table below), some variables like hemo and bloodureanitro show clear, significant monotonic relationships with LOS. Others (like sodium, glucose, bmi, etc.) show very low or no correlation in their original form. However, as previously discussed, we acknowledged that biological variables often present in different scales, magnitudes, and distributions to what we immediately need.

Acknowledging this brought us to realize that variables with low correlation can still be significant after transformation. They may indeed show importance when either transformed (log, square, inverse) or interacted with other variables. How these variables have been transformed has been discussed in the methods for objective 4.

Spearman_Correlation__Ordered_Variables_

	Spearman Corr	p-value
hemo	0.2141585950180550	0.0
hematocrit	-0.05759544436981500	3.08401027509797E-74
neutrophils	-0.09114654708030550	1.9611906254314E-183
sodium	-0.0025791071021070600	0.4147424297219220
bloodureanitro	0.11720669099117700	9.26255994415106E-303
creatinine	-0.0029931925056080000	0.3438829838984330
bmi	-0.00013445044746306900	0.9660870116619550
pulse	0.004187954194033680	0.18539126324030700
respiration	0.001390000252546180	0.6602620522638500
glucose	-0.00015614803423965400	0.9606182777025790

For the categorical variables, as observable in table below, all variables fall into the small effect range (0.01–0.06). Furthermore, values in the range of 0.03-0.04 (e.g., irondef, malnutrition) suggest small yet potentially meaningful contributions, especially in large datasets like ours ($n = 100,000$). In conclusion, none of the categorical variables have a large effect, yet small, but notable correlations with Length of Stay. These small effects will still be meaningful in combination with the other variables in regression modeling, as we will see in the results for objective 4.

Correlation_Ratio__Categorical_Variables_

	Correlation Ratio	ANOVA p-value
malnutrition	0.030414233816531900	0.0
dialysisrenalendstage	0.028796328149468100	0.0
irondef	0.03757302799538760	0.0
pneum	0.01834972195286710	0.0
substancedependence	0.02187023246978980	0.0
psychologicaldisordermajor	0.08221064895083310	0.0
depress	0.01473985263454700	0.0
psychother	0.036747146622564700	0.0

The findings from Objective 3 offer critical insight into how internal patient variables relate to hospital Length of Stay (LOS). While some variables, such as hemo and bloodureanitro, displayed clear monotonic correlations with LOS, others initially appeared unrelated. However, this analysis emphasized that low raw correlation does not imply irrelevance; many variables demonstrated significant predictive value only after appropriate transformations or interactions were applied.

These results form a strong foundation for the next phase of our research, where we focus on predictive modeling. Understanding which variables carry meaningful relationships with LOS, whether directly or through transformed variables, helps us make informed decisions about feature selection. Beyond modeling, these insights also support hospital management by

highlighting key clinical indicators that can be used to anticipate patient stays, allocate resources more effectively, and improve overall care planning.

Objective 4

The fourth objective of our project was the development of a predictive model for hospital Length of Stay (LOS), using internal physiological and clinical variables identified in the earlier stages of analysis. Drawing on the insights from our correlation and transformation work, we constructed a dataset optimized for modeling, that includes not only the original body levels variables, but also meaningful transformations (e.g., squared, inverse, and log-transformed variables) and interaction terms that had demonstrated significance in regression testing. The aim was to move beyond simple associations and create a model with real predictive power that can be practically applied in a hospital setting.

Using this enriched dataset, we trained a model designed to estimate LOS based on a patient's internal metrics at the time of admission. The model was built with the goal of assisting healthcare institutions in anticipating the length of stay of each patient. Importantly, only variables that were clinically relevant and statistically validated were retained in the final model to ensure both accuracy and interpretability. In the following section, we present the model's performance metrics, examine the relative importance of each predictor, and discuss the practical implications of using such a tool for hospital operations.

Our findings

Starting from the model performance indicators, the regression model achieved an R-squared of 0.266, meaning it explains approximately 26.6% of the variation in patients' Length of Stay (LOS). While this may seem modest at first, it is actually a strong result in the context of healthcare data, where patient outcomes like LOS are influenced by a wide range of factors, many of which are unmeasured, external (opposed to objectives 1 and 2), or behavioral. In clinical and administrative modeling, R^2 values between 0.20 and 0.30 are common and often considered strong for operational planning and forecasting.

The adjusted R-squared is also 0.266, indicating that the model includes a good balance of explanatory variables without overfitting, even with a relatively large number of predictors (29 in total). This suggests that the transformations and interaction terms included in the model are not just statistically significant, they also contribute meaningfully to the overall predictive power.

Furthermore, the model's F-statistic is 1248.9 with a p-value well below 0.001, confirming that the model as a whole is statistically significant. In other words, the likelihood that these results occurred by chance is effectively zero. With 100,000 observations, the model benefits from high statistical power, which makes even subtle effects detectable and strengthens the reliability of the estimates.

In summary, this model is both statistically robust and practically informative, offering meaningful predictive insights that can be used for hospital planning, resource allocation, and identifying risk factors for extended hospitalization.

R-squared: 0.266, Adjusted R-squared: 0.266
F-statistic: 1248.901 df(29,99970), p.value < .001
Nr obs: 100,000

Moving to the coefficient analysis, most predictors have very low p-values (< 0.001), many have large t-values (often >10 or even >40), and the model includes all selected transformed and interaction terms. These features strongly suggest that the model has excellent explanatory power and is statistically robust.

	coefficient	std.error	t.value	p.value
(Intercept)	416.973	9.585	43.501	< .001 ***
dialysisrenalendstage	0.605	0.037	16.195	< .001 ***
irondef	0.679	0.023	28.933	< .001 ***
pneum	-0.728	0.242	-3.007	0.003 **
substancedependence	0.860	0.027	32.070	< .001 ***
psychologicaldisordermajor	1.189	0.017	70.842	< .001 ***
depress	1.051	0.060	17.382	< .001 ***
psychother	0.671	0.034	19.842	< .001 ***
malnutrition	0.493	0.033	14.997	< .001 ***
hemo	1.027	0.035	29.667	< .001 ***
hematocrit	0.541	0.011	48.230	< .001 ***
neutrophils	0.040	0.002	19.229	< .001 ***
sodium	-5.678	0.138	-41.117	< .001 ***
bloodureanitro	0.008	0.001	12.913	< .001 ***
creatinine	7.867	0.224	35.056	< .001 ***
bmi	-2.660	0.067	-39.560	< .001 ***
pulse	0.074	0.002	34.100	< .001 ***
respiration	-3.140	0.084	-37.394	< .001 ***
inv_hematocrit	75.471	1.531	49.294	< .001 ***
log_neutrophils	-0.411	0.028	-14.772	< .001 ***
glucose	-0.056	0.001	-39.154	< .001 ***
glucose_sq	0.000	0.000	39.453	< .001 ***
log_creatinine	-8.379	0.235	-35.588	< .001 ***
inv_pulse	362.960	10.425	34.818	< .001 ***
respiration_sq	0.247	0.006	38.024	< .001 ***
respiration_pneum	0.212	0.038	5.631	< .001 ***
bmi_2	0.045	0.001	39.609	< .001 ***
sodium_2	0.021	0.001	41.101	< .001 ***
depress_psych	-0.773	0.070	-11.038	< .001 ***
hemo_bun_interaction	-0.006	0.001	-4.820	< .001 ***

Following is a comprehensive summary of each variable, its coefficient, and the relative interpretation when used in a predictive model for the LOS.

Variable	Coefficient	Interpretation
(Intercept)	416.973	Baseline LOS when all predictors = 0; not directly interpretable in clinical terms.
dialysisrenalendstage	0.605	End-stage renal disease adds to LOS — likely due to complex care needs.
irondef	0.679	Iron deficiency is associated with longer stays, possibly reflecting poor baseline health.
pneum	-0.728	Pneumonia associated with shorter LOS — may suggest rapid treatment due to urgency.

substancedependence	0.86	Strongly associated with longer stays, likely due to complications or social factors.
psychologicaldisordermajor	1.189	Major psychiatric disorders significantly increase LOS — consistent with clinical expectations.
depress	1.051	Depression independently increases LOS — possibly due to slower recovery or behavioral barriers.
psychother	0.671	Psychological therapy linked to longer LOS, possibly due to ongoing mental health support during admission.
malnutrition	0.493	Malnutrition increases LOS — common in frail or elderly patients.
hemo	1.027	Lower hemoglobin linked to longer stays; anemia complicates recovery.
hematocrit	0.541	Lower hematocrit (RBC % volume) extends LOS.
neutrophils	0.04	Elevated neutrophils (inflammation) slightly increase LOS.
sodium	-5.678	Higher sodium shortens LOS — very low sodium likely prolongs stay due to complications.
bloodureanitro	0.008	Higher BUN slightly increases LOS — linked to kidney function.
creatinine	7.867	High raw creatinine increases LOS — indicator of renal dysfunction.
bmi	-2.66	Higher BMI associated with shorter LOS — potentially reflects underweight risk.
pulse	0.074	Higher pulse slightly increases LOS — may indicate clinical instability.
respiration	-3.14	Higher respiratory rate associated with shorter LOS — possibly early resolution or triage effect.
inv_hematocrit	75.471	Low hematocrit dramatically increases LOS.
log_neutrophils	-0.411	High neutrophils (log scale) slightly decrease LOS — may reflect resolving infection.
glucose	-0.056	Higher glucose linked to marginally shorter stays.
glucose_sq	0.0	Nonlinear effect — both very high and low glucose may increase LOS.
log_creatinine	-8.379	Strong negative effect — counterbalances raw creatinine.
inv_pulse	362.96	Low pulse dramatically increases LOS — possible indicator of frailty or critical illness.
respiration_sq	0.247	Extreme respiration values (high or low) increase LOS.
respiration_pneum	0.212	Patients with pneumonia + abnormal respiration stay longer (respiratory conditions force prolonged stay)
bmi_2	0.045	BMI ² confirms nonlinear U-shaped effect: both low and high BMI are problematic.
sodium_2	0.021	Sodium ² supports a U-shaped relationship — extremes increase LOS.
depress_psych	-0.773	Negative interaction — co-occurrence reduces marginal impact (possible ceiling effect).
hemo_bun_interaction	-0.006	Slightly negative interaction — anemia + high BUN co-occurrence adds less LOS than expected.

The results of Objective 4 mark a significant milestone in our project: the successful creation of a predictive model capable of estimating hospital Length of Stay (LOS) using a refined set of internal physiological and clinical variables. By incorporating not only raw indicators but also carefully selected transformations and interaction terms, we were able to capture complex, nonlinear relationships that improved the model's accuracy and clinical relevance. The model's strong statistical validity, supported by low p-values and robust sample size, confirms its potential as a practical forecasting tool for hospitals.

These findings serve as a critical foundation for the next phase of our work. With the model now built and validated, we are prepared to move forward to Objective 5, where we will translate the model's insights into actionable recommendations for hospitals. By applying the predictive capabilities of the model, we will aim to suggest targeted changes, such as improved resource planning, risk-based patient management, and proactive care strategies, that can lead to increased hospital efficiency and better patient outcomes.