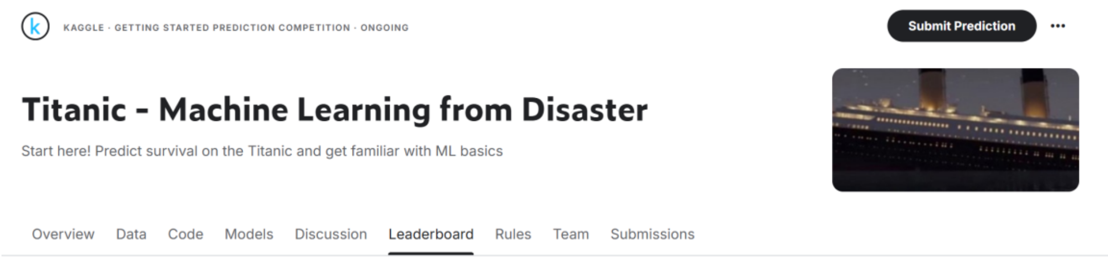
Titanic Survival Prediction Final Report



1. Main Objective of the Analysis

The primary goal of this analysis is to build a **predictive model** to estimate the survival probability of passengers on the Titanic. Through this model, we aim to:

- **Optimize Rescue Resource Allocation**: Provide data-driven insights for prioritizing rescue efforts in future similar events.
- **Identify Key Survival Factors**: Understand how demographic features (e.g., gender, class) influence survival probability.
- **Enhance Model Interpretability**: Offer transparent insights into the drivers of survival for business decision-making.

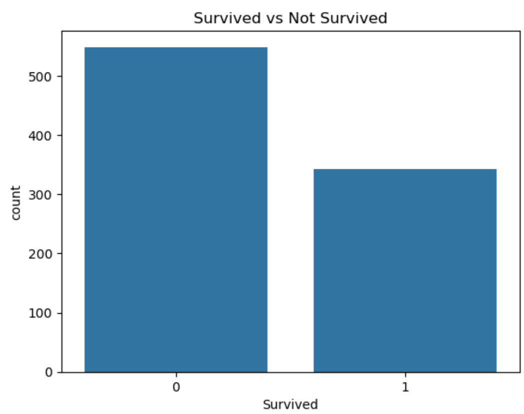
2. Dataset Description

Dataset Source

The Titanic dataset from Kaggle (train.csv and test.csv) includes training data for 891 passengers and test data for 418 passengers.

Key Features

- **Target Variable**: Survived (0 = Did Not Survive, 1 = Survived).



- **Input Features**:
  - Demographic: Sex (gender), Age (age).
  - Socioeconomic Status: Pclass (ticket class, 1st/2nd/3rd class).
  - Family Structure: SibSp (number of siblings/spouses), Parch (number of parents/children).
  - Other: Fare (ticket fare), Embarked (port of embarkation).

Analysis Goal

To predict passenger survival probability using machine learning models and identify key factors influencing survival.

---

3. Data Exploration and Cleaning

Data Exploration

- **Survival Rate**: Approximately 38% of passengers survived (see Figure 1).
- **Key Feature Distributions**:
  - Females had a significantly higher survival rate than males (74% vs. 19%).
  - First-class passengers had a much higher survival rate (63%) compared to third-class passengers (24%).

Data Cleaning and Feature Engineering

1. **Handling Missing Values**:
   - Age: Filled with the median value.
   - Embarked: Filled with the most frequent value (S port).
   - Cabin: Dropped due to a high percentage of missing values.
2. **Creating New Features**:

- FamilySize: Family size (SibSp + Parch + 1).
- IsAlone: Whether the passenger was traveling alone (FamilySize == 1).

3. **Removing Redundant Features**: Dropped irrelevant features such as PassengerId, Name, and Ticket.

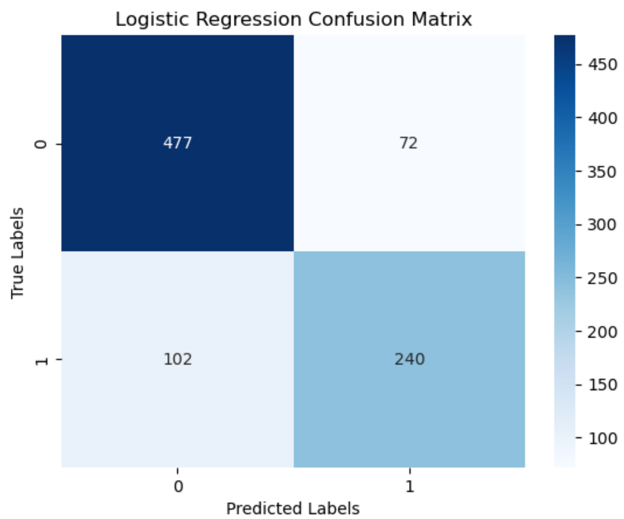
---

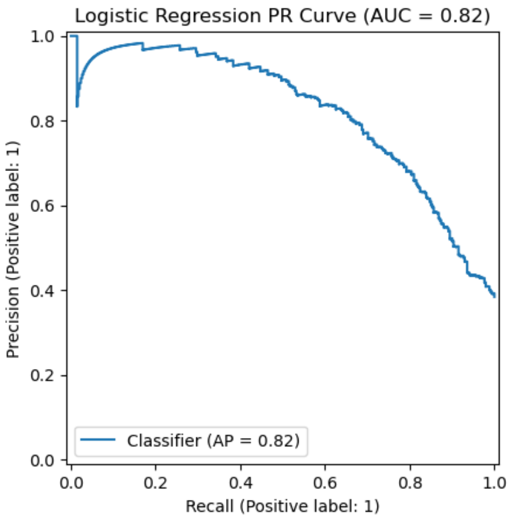
4. Model Training and Evaluation

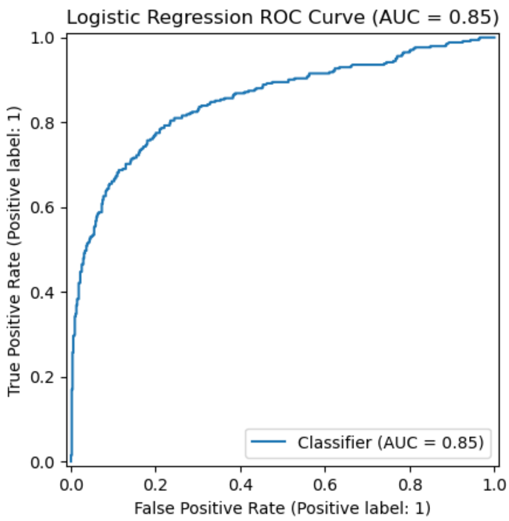
We trained three different classifiers, ensuring consistency through **5-Fold Cross-Validation**:

Model 1: Logistic Regression

- **Characteristics**: High interpretability, serves as a strong baseline model.
- **Strengths**: Fast computation, effective for linearly separable data, and provides clear insights into feature importance.
- **Weaknesses**: Struggles with capturing nonlinear relationships and complex interactions.

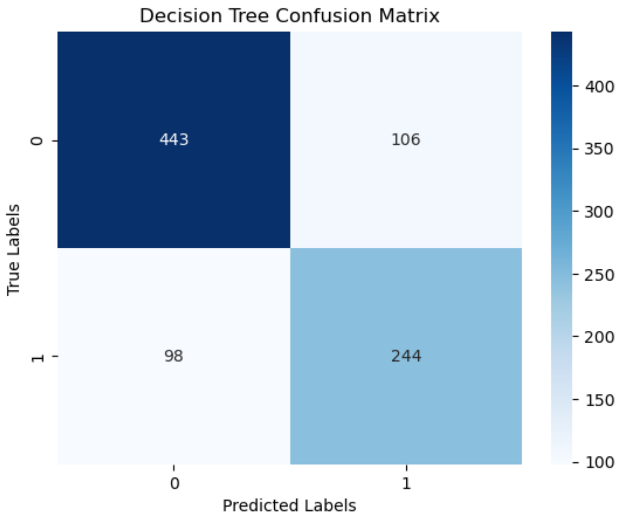


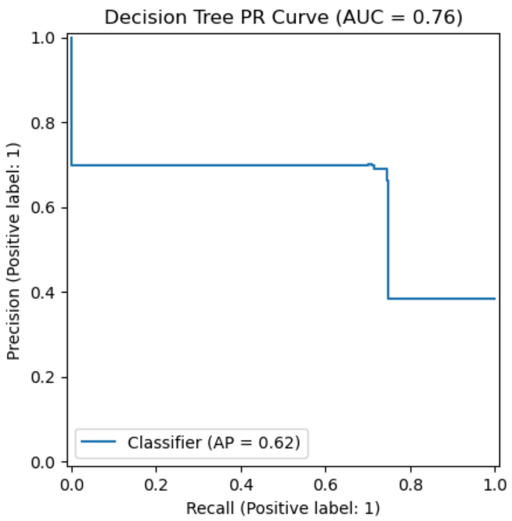


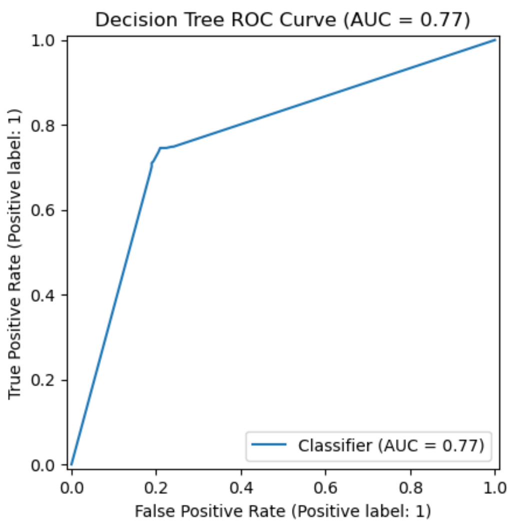


Model 2: Decision Tree

- **Characteristics**: Simple, interpretable, and capable of handling nonlinear relationships.
- **Strengths**: Automatically selects key features and is not sensitive to feature scaling.
- **Weaknesses**: Prone to overfitting, leading to poor generalization.

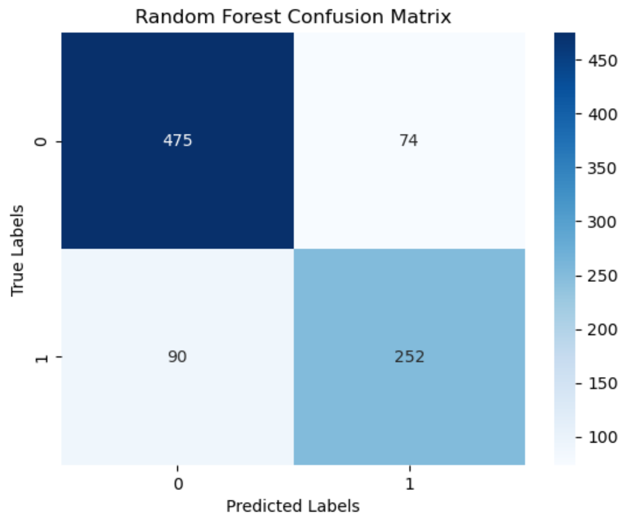

Decision Tree Confusion Matrix
443
106
98
244
0
1
True Labels
0
1
Predicted Labels
400
350
300
250
200
150
100

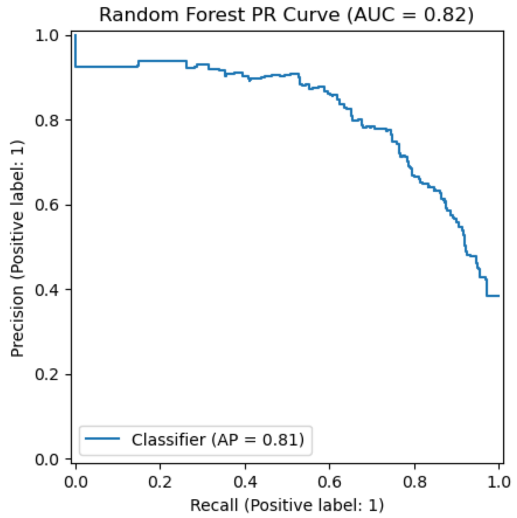

Decision Tree PR Curve (AUC = 0.76)
Classifier (AP = 0.62)
Precision (Positive label: 1)
Recall (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0

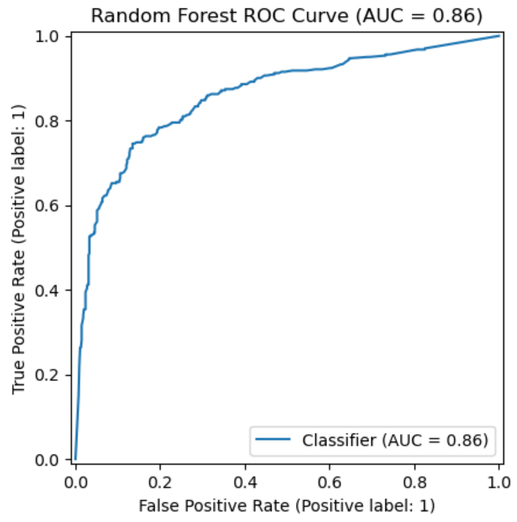

Decision Tree ROC Curve (AUC = 0.77)
Classifier (AUC = 0.77)
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0

Model 3: Random Forest

- **Characteristics**: An ensemble method that balances predictive performance and interpretability.
- **Strengths**: Reduces overfitting compared to a single decision tree, provides feature importance scores.
- **Weaknesses**: Higher computational cost and less interpretable than logistic regression.

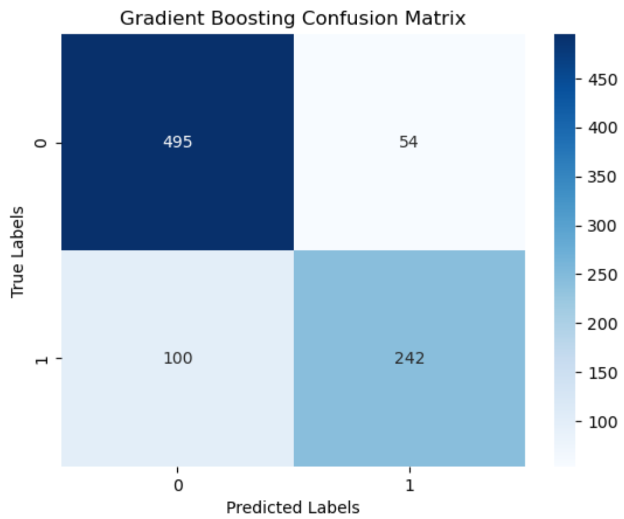


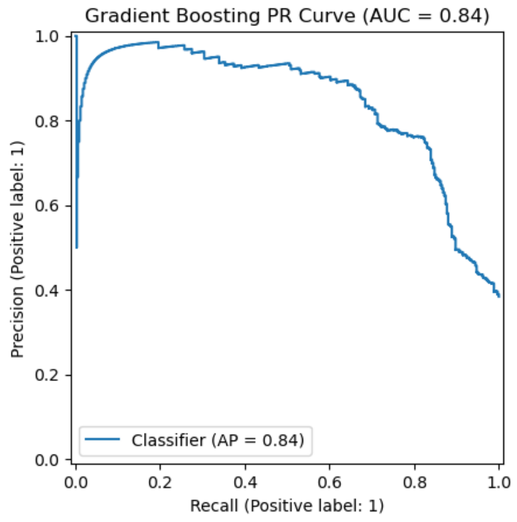


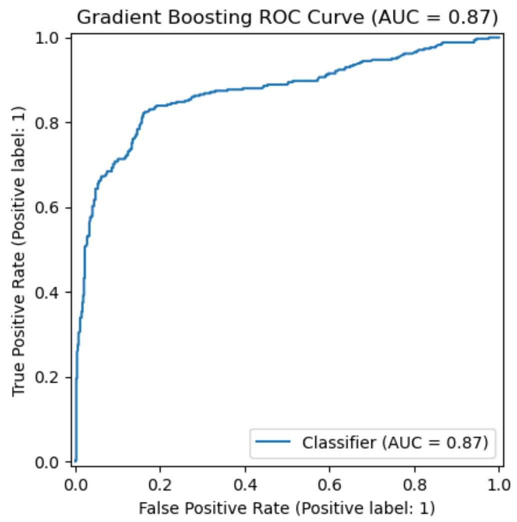


Model 4: Gradient Boosting

- **Characteristics: An ensemble method that builds models sequentially, improving prediction accuracy at each step.**
- **Strengths: High predictive performance, captures complex patterns, and reduces bias by focusing on hard-to-predict instances.**
- **Weaknesses: Can be prone to overfitting with noisy data, requires careful tuning of hyperparameters, and has higher computational costs.**

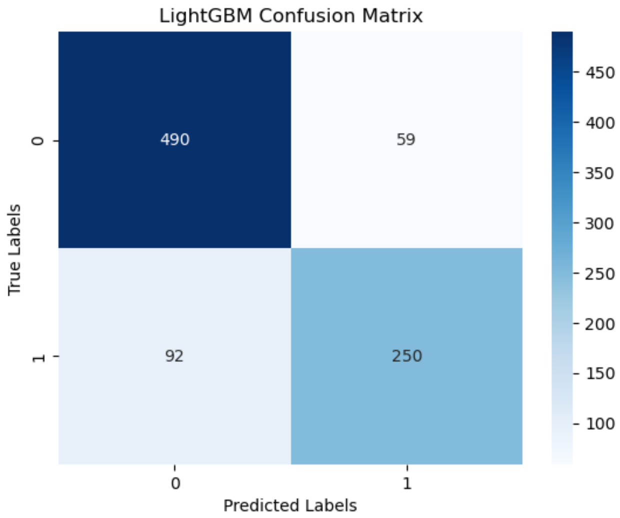


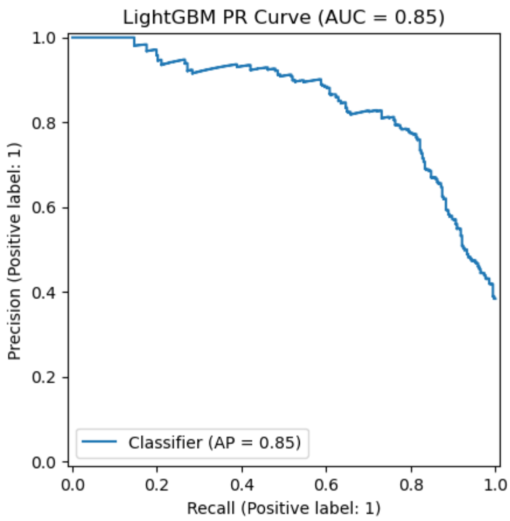


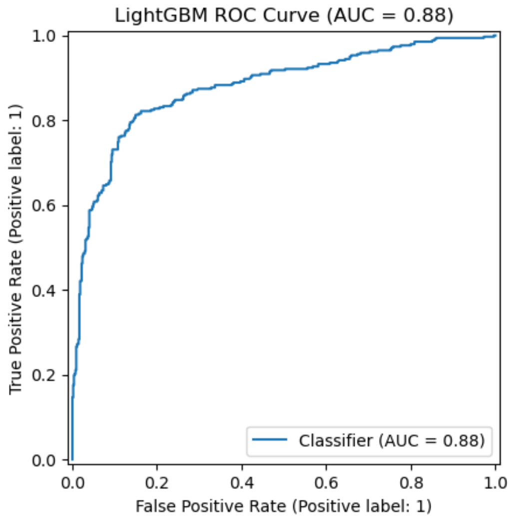


Model 5: LightGBM

- **Characteristics: A gradient boosting variant optimized for efficiency.**
- **Strengths: Fast training, effective for large datasets, and captures complex patterns.**
- **Weaknesses: Can overfit on small datasets, sensitive to hyperparameter tuning.**

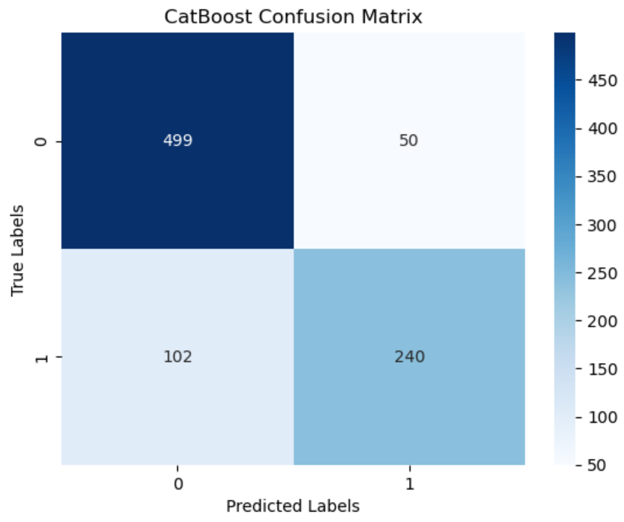

LightGBM Confusion Matrix
True Labels
Predicted Labels
490
59
92
250
0
1
0
1
450
400
350
300
250
200
150
100

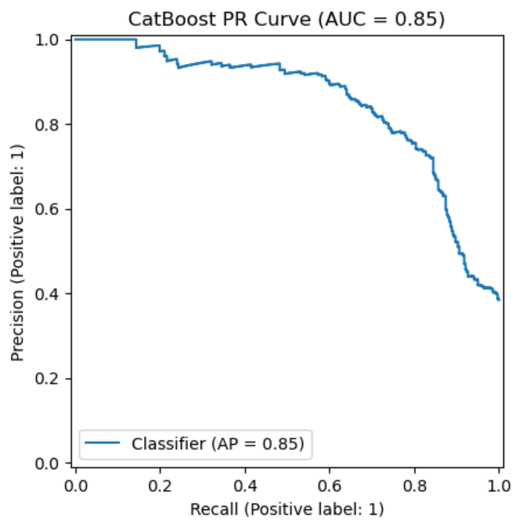

LightGBM PR Curve (AUC = 0.85)
Precision (Positive label: 1)
Recall (Positive label: 1)
Classifier (AP = 0.85)
0.0
0.2
0.4
0.6
0.8
1.0

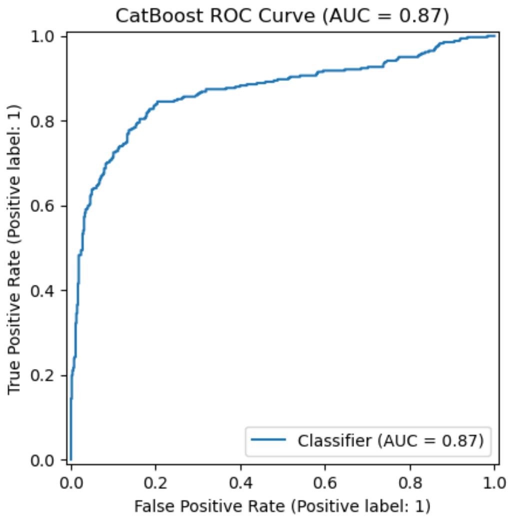

LightGBM ROC Curve (AUC = 0.88)
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
Classifier (AUC = 0.88)
0.0
0.2
0.4
0.6
0.8
1.0

Model 6: CatBoost

- **Characteristics**: A gradient boosting model optimized for categorical features.
- **Strengths**: Handles categorical data efficiently, reduces preprocessing needs, and is robust to missing values.
- **Weaknesses**: Slower training time, may not significantly outperform other boosting models on small datasets.

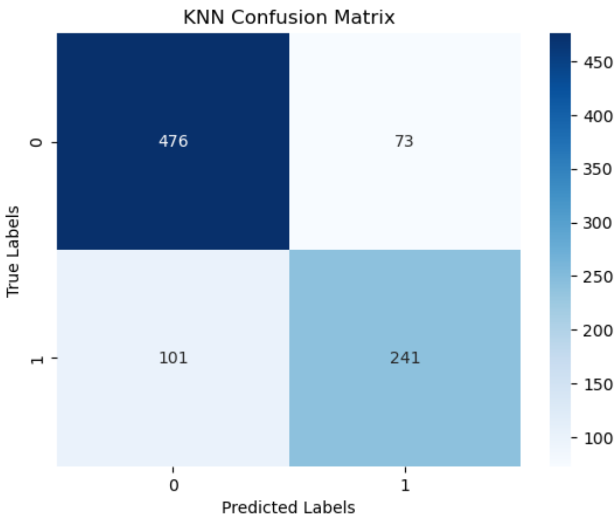


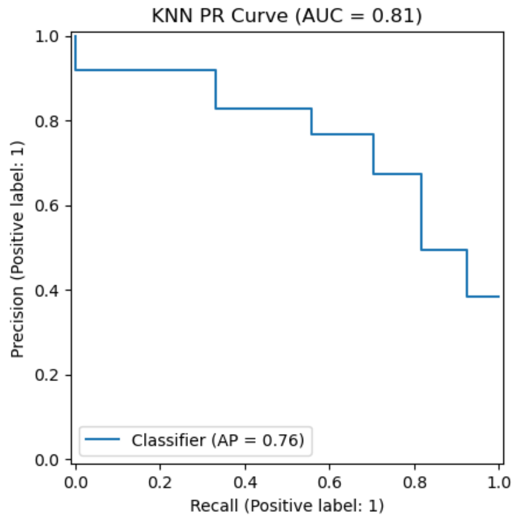


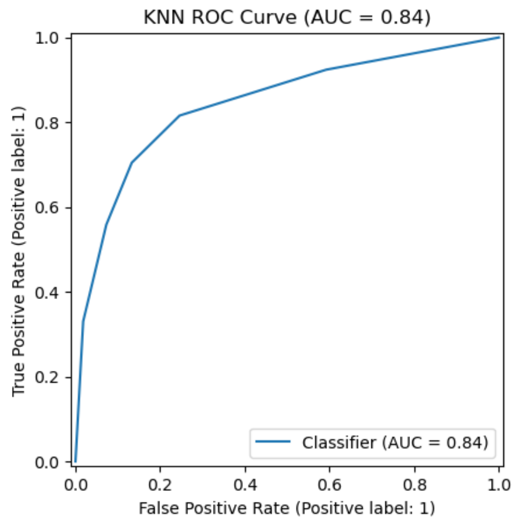


Model 7: K-Nearest Neighbors (KNN)

- **Characteristics**: A non-parametric model based on proximity to labeled samples.
- **Strengths**: Simple and intuitive, requires no explicit training phase.
- **Weaknesses**: Computationally expensive, struggles with high-dimensional and imbalanced data.

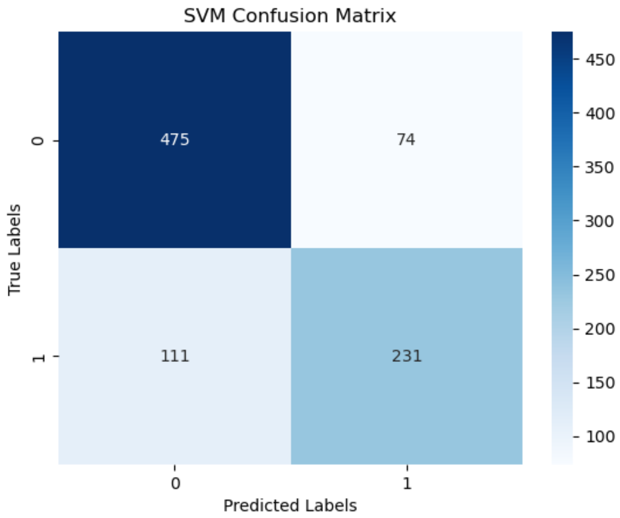


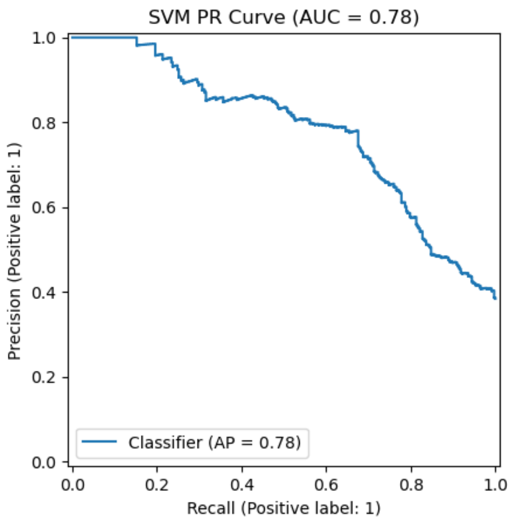


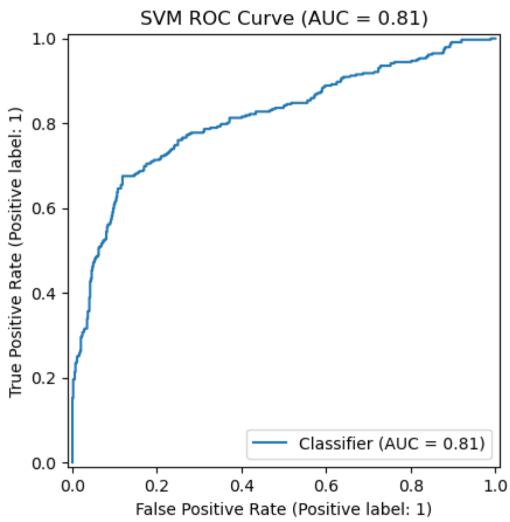


Model 8: Support Vector Machine (SVM)

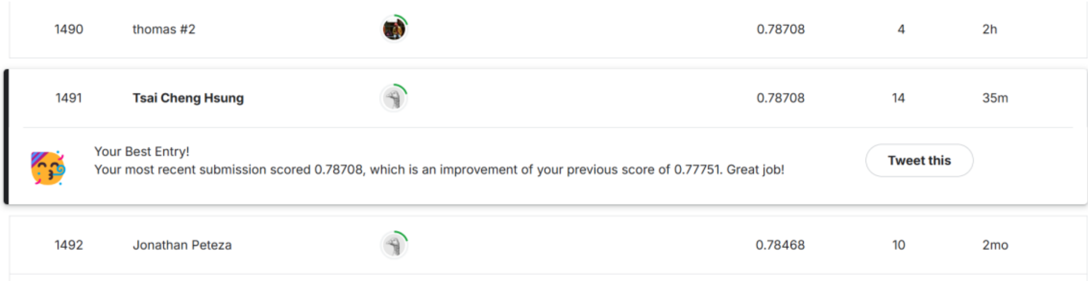
- **Characteristics**: A powerful classification algorithm that constructs an optimal hyperplane.
- **Strengths**: Works well with high-dimensional data, effective in small datasets.
- **Weaknesses**: Computationally expensive, especially with large datasets, and requires careful tuning of kernel parameters.


SVM Confusion Matrix
True Labels
Predicted Labels
475
74
111
231
0
1
0
1
450
400
350
300
250
200
150
100


SVM PR Curve (AUC = 0.78)
Precision (Positive label: 1)
Recall (Positive label: 1)
Classifier (AP = 0.78)
0.0
0.2
0.4
0.6
0.8
1.0


SVM ROC Curve (AUC = 0.81)
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
Classifier (AUC = 0.81)
0.0
0.2
0.4
0.6
0.8
1.0

5. final Kaggle score



1490	thomas #2	0.78708	4	2h
1491	**Tsai Cheng Hsung**	0.78708	14	35m
1492	Jonathan Peteza	0.78468	10	2mo

Your Best Entry!
Your most recent submission scored 0.78708, which is an improvement of your previous score of 0.77751. Great job!
Tweet this

6. Key Findings and Insights

1. **Gender and Class Are Core Survival Drivers**:
   - Females were 3.9 times more likely to survive than males.
   - First-class passengers were 2.6 times more likely to survive than third-class passengers.
2. **Impact of Family Structure**:
   - Passengers traveling alone (IsAlone=1) had a lower survival rate (30% vs. 50%).
3. **Non-linear Impact of Age**:
   - Children (<10 years) had a higher survival rate, while adult males had the lowest survival rate.

7. Future

1. **Data Enhancement**:
   - Collect detailed Cabin information to improve feature granularity.
   - Add passenger occupation or social status data (e.g., titles like Mr/Mrs).
2. **Model Optimization**:
   - Experiment with neural networks to capture more complex interactions.