Lung Cancer Survival Prediction using Machine Learning

Shravani Bande

1 Introduction

Lung cancer is one of the leading causes of cancer-related deaths globally. Predicting the survival of a lung cancer patient can help in better treatment planning and early interventions. This project leverages machine learning models to predict the likelihood of a patient's survival based on medical history and diagnosis data.

2 Dataset Description

The dataset consists of 890,000 records with features such as:

- Demographics (age, gender, country)
- Medical conditions (asthma, cirrhosis, hypertension, etc.)
- Diagnosis and treatment-related data
- Target column: `survived` (0 = No, 1 = Yes)

3 Data Preprocessing

- Handled missing and infinite values.
- Converted object columns (e.g., country, cancer stage) using one-hot encoding.
- Calculated `treatment_duration` from treatment start and end dates.
- Normalized numerical columns for better model performance.

4 Exploratory Data Analysis

Explored correlations and data imbalance:

- 77.9% patients did not survive.
- Certain features like BMI, age, and cancer stage showed strong correlations with survival.

5 Model Training

We trained multiple models including:

- Logistic Regression
- Random Forest
- XGBoost (best performing)

XGBoost achieved an accuracy of **77.7%**. Due to data imbalance, recall for class 1 (survived) was lower.

6 Prediction

A new patient's survival is predicted using the trained model:

```
Predicted:  Did not survive  Actual:  Did not survive
```

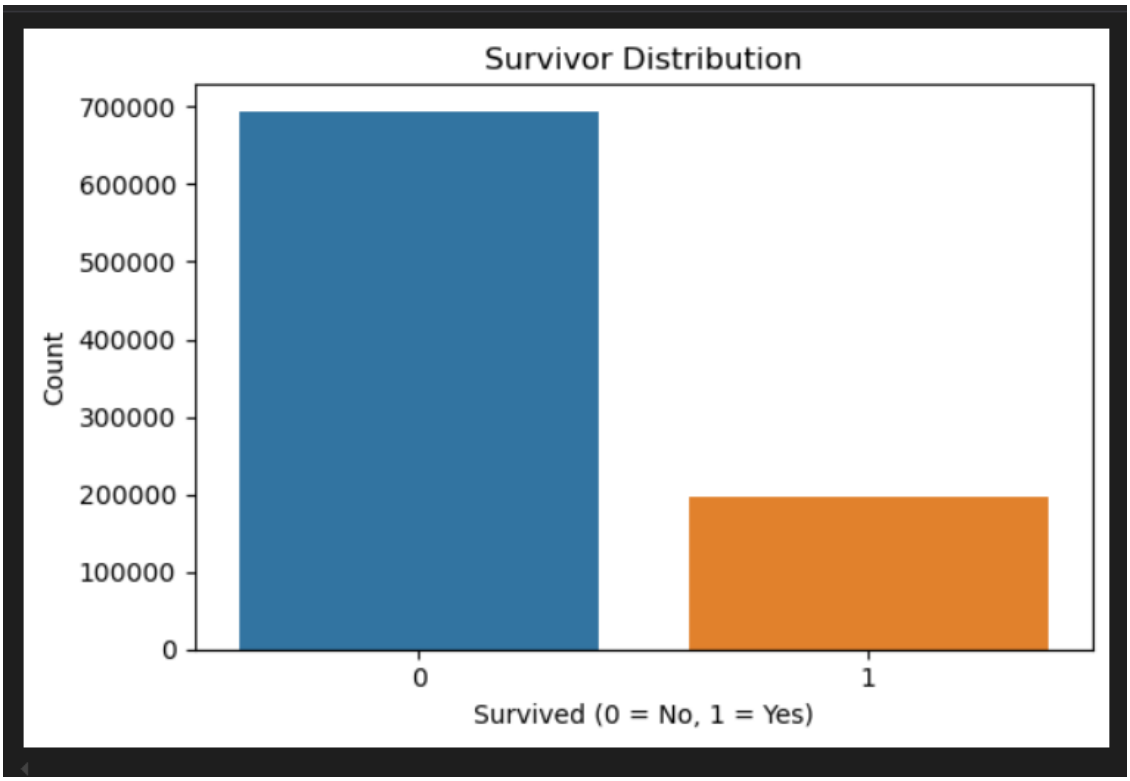
The model takes a dictionary of patient data, applies preprocessing, and returns the survival prediction.

7 Code Architecture

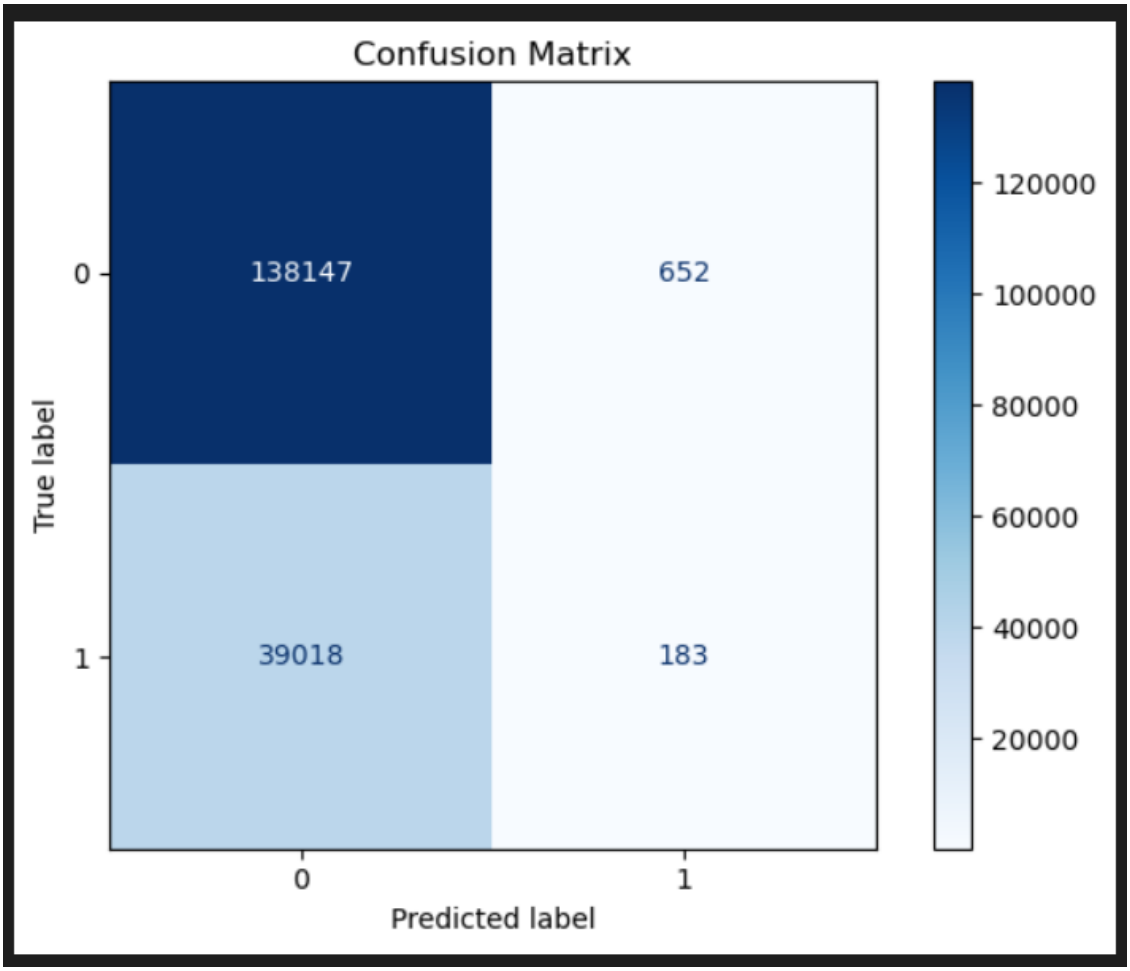
- `eda.ipynb` – For data exploration and cleaning.
- `model.py` – Contains training and evaluation logic.

- `predict.py` – Loads trained model and runs prediction.
- `data/` – Dataset used for training.
- `models/` – Trained model saved as a .pkl file.
- `requirements.txt` – Python packages used.
- `README.md` – Summary of the project for GitHub.

8 Visualizations and Screenshots



Survivor Distribution



Confusion Matrix of Final Model

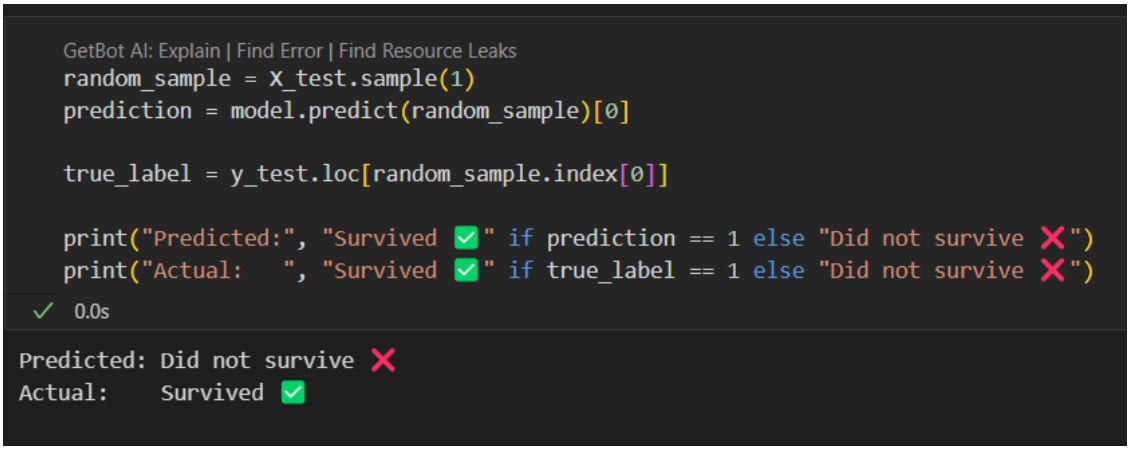
```
GetBot AI: Explain | Find Error | Find Resource Leaks
random_sample = X_test.sample(1)
prediction = model.predict(random_sample)[0]

true_label = y_test.loc[random_sample.index[0]]

print("Predicted:", "Survived ✅" if prediction == 1 else "Did not survive ❌")
print("Actual:   ", "Survived ✅" if true_label == 1 else "Did not survive ❌")
```

```
Predicted: Did not survive ❌
Actual:    Did not survive ❌
```

Sample Patient Prediction Output



Sample Patient Prediction Output

9 Conclusion

This project demonstrates how machine learning can be applied to medical data to predict patient survival. While accuracy is high, handling imbalanced data and improving recall for minority class remain future goals.

10 References

- Scikit-learn Documentation: `https://scikit-learn.org`
- XGBoost Documentation: `https://xgboost.readthedocs.io`
- Dataset Source (e.g., Kaggle or project-provided)