FINAL PROJECT: LUNG CANCER PREDICTION

Luke Profio

Project Overview

- Early detection of lung cancer dramatically increases survival rates, but diagnosis is often delayed until advanced stages.
- In this project, we build a supervised learning pipeline that predicts whether a patient is likely to have lung cancer from an anonymous survey covering demographics, lifestyle habits and respiratory-related symptoms.

Objectives

1. Clean, explore and visualize the survey data.
2. Build and evaluate predictive models (Logistic Regression as baseline; Random Forest as ensemble)
3. Interpret which features drive the predictions.
4. Summarize findings and outline limitations of the study.

About the Dataset

- This is a publicly-available Kaggle dataset of hypothetical lung cancer patients that was synthetically generated.
  - **Nelson, S. G.** (2023). *Lung Cancer Prediction* [Kaggle Notebook]. Kaggle. Retrieved July 15, 2025, from https://www.kaggle.com/code/sandragracenelson/lung-cancer-prediction/notebook
- This contains 11kB of tabulated data, with over 310 records and 16 features (14 numeric, 2 non-numeric). Outside of demographic features such as gender and age, there are a variety of symptom-based features such as anxiety and allergy, in addition to the target feature, LUNG_CANCER which is converted into a numeric value.

Data Cleaning

- Missing values were checked for, while duplicates and outliers were not relevant for this dataset given it didn’t have an identifier, and all numeric columns were re-encoded as 1’s/0’s.
- The target variable was mapped from “YES/NO” to 1/0 to provide the numeric form needed by classifiers.
- All 2’s were replaced with 1’s, and all 1’s were replaced with 0’s (where 2 = true/positive and 1 = false/negative) across non-object columns to ensure consistency in Boolean features, and avoid unintended ordinal interpretations.
- Gender was re-encoded from M/F to 1/0 to ensure a uniform numeric representation of the data for downstream analysis.

Data Cleaning

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   GENDER                 309 non-null    int64
 1   AGE                    309 non-null    int64
 2   SMOKING                309 non-null    int64
 3   YELLOW_FINGERS         309 non-null    int64
 4   ANXIETY                309 non-null    int64
 5   PEER_PRESSURE          309 non-null    int64
 6   CHRONIC_DISEASE        309 non-null    int64
 7   FATIGUE                309 non-null    int64
 8   ALLERGY                309 non-null    int64
 9   WHEEZING               309 non-null    int64
 10  ALCOHOL_CONSUMING      309 non-null    int64
 11  COUGHING               309 non-null    int64
 12  SHORTNESS_OF_BREATH    309 non-null    int64
 13  SWALLOWING_DIFFICULTY  309 non-null    int64
 14  CHEST_PAIN             309 non-null    int64
 15  LUNG_CANCER            309 non-null    int64
dtypes: int64(16)
memory usage: 38.8 KB
```

	missing
GENDER	0
AGE	0
SMOKING	0
YELLOW_FINGERS	0
ANXIETY	0
PEER_PRESSURE	0
CHRONIC_DISEASE	0
FATIGUE	0
ALLERGY	0
WHEEZING	0
ALCOHOL_CONSUMING	0
COUGHING	0
SHORTNESS_OF_BREATH	0
SWALLOWING_DIFFICULTY	0
CHEST_PAIN	0
LUNG_CANCER	0

GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC_DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL_CONSUMING	COUGHING	SHORTNESS_OF_BREATH	SWALLOWING_DIFFICULTY	CHEST_PAIN	LUNG_CANCER
1	69	0	1	1	0	0	1	0	1	1	1	1	1	1	1
1	74	1	0	0	0	1	1	1	0	0	0	1	1	1	1
0	59	0	0	0	1	0	1	0	1	0	1	1	0	1	0
1	63	1	1	1	0	0	0	0	0	1	0	0	1	1	0
0	63	0	1	0	0	0	0	0	1	0	1	1	0	0	0

Data Cleaning – Conclusions

- There were no missing values found.
- There weren’t any major issues with data cleaning.
  - This is likely due to the dataset being synthetic. In real-world, scaled datasets there would be outliers, duplicates, and missing values to be cleaned.
- Overall, the strategy involved remapping the dataset into 0’s and 1’s which were logical and compatible with the analytic modelling to be performed.

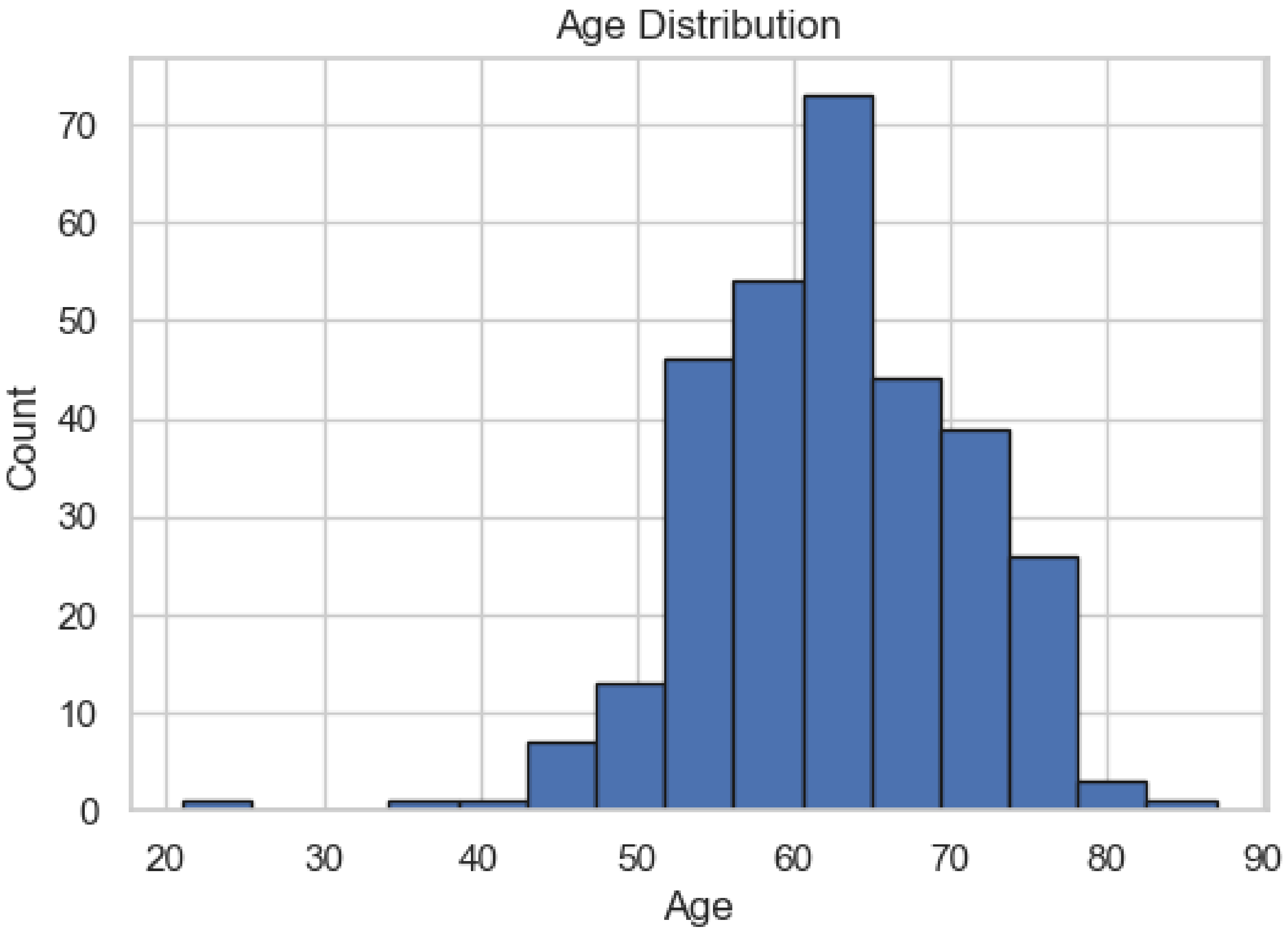
Exploratory Data Analysis

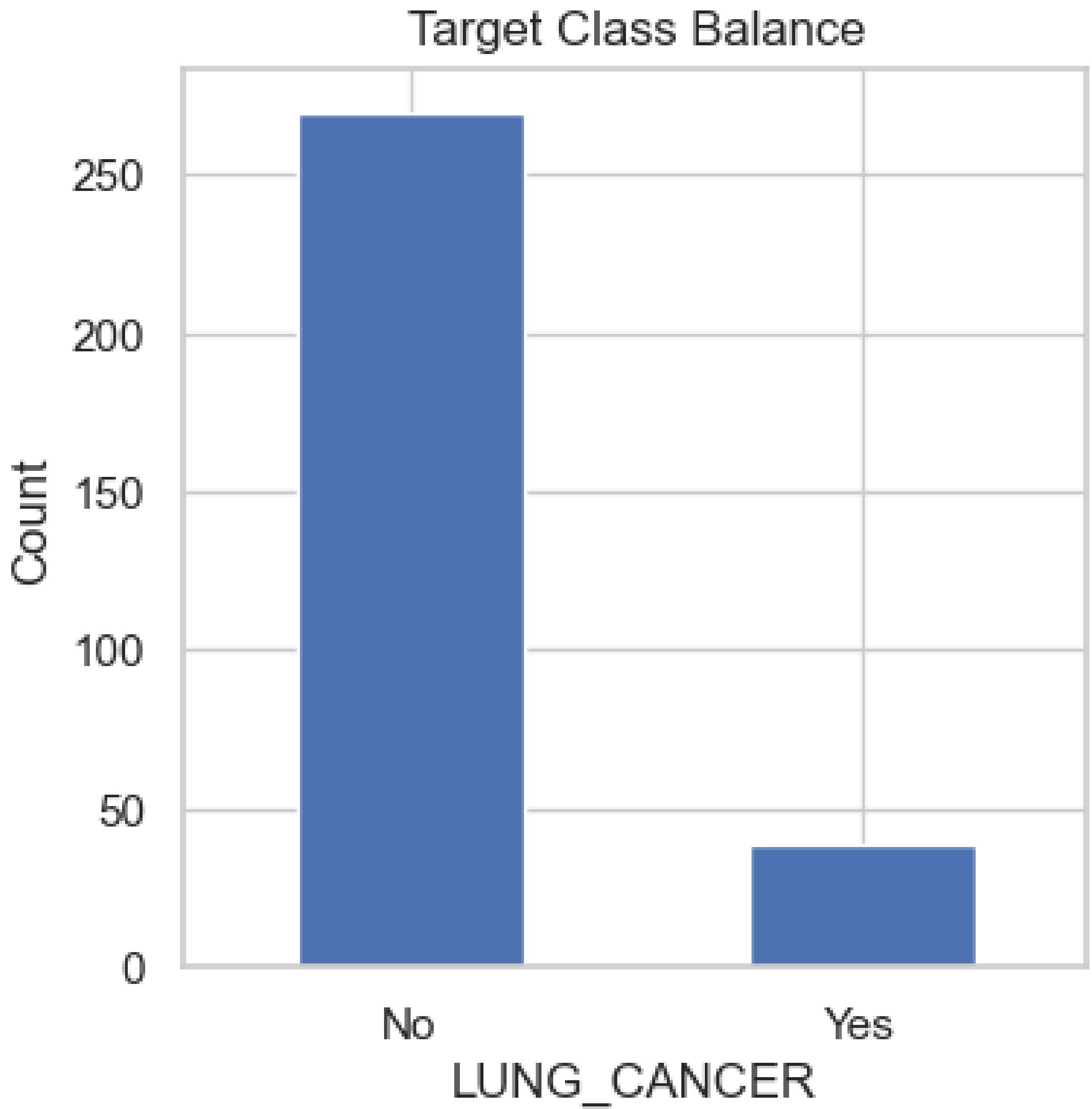
- The purpose of EDA with this dataset was to identify bias, skewness, multicollinearity, and other factors that may impact the subsequent analysis.
- Plotted a histogram of age and each target variable to identify potential bias in the dataset.
- Generated a heatmap of pairwise correlations among targets to assess multicollinearity and strong associations.
- Tree-based importance scores were used to rank predictors, helping to guide feature selection.

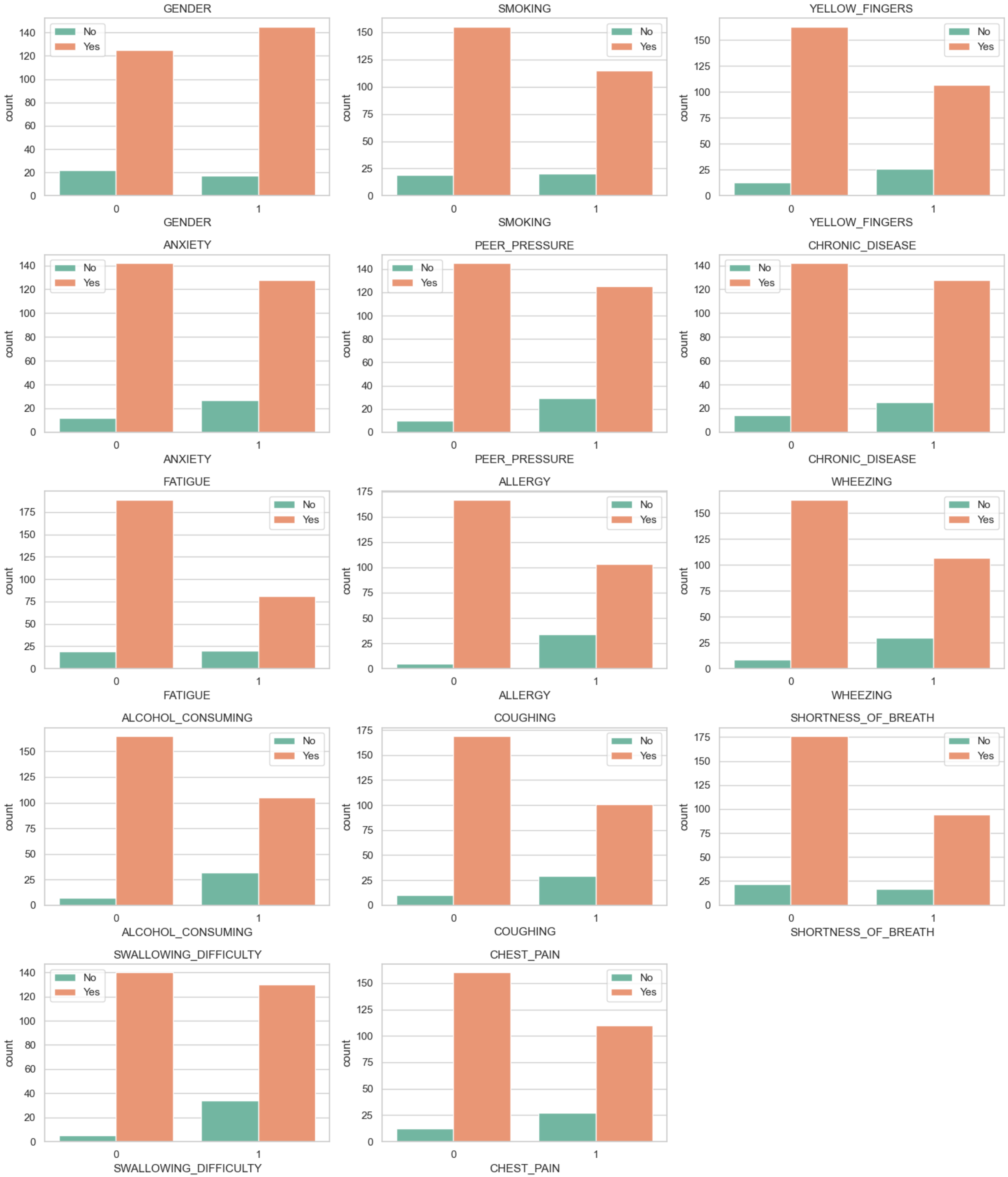
Exploratory Data Analysis - Conclusions

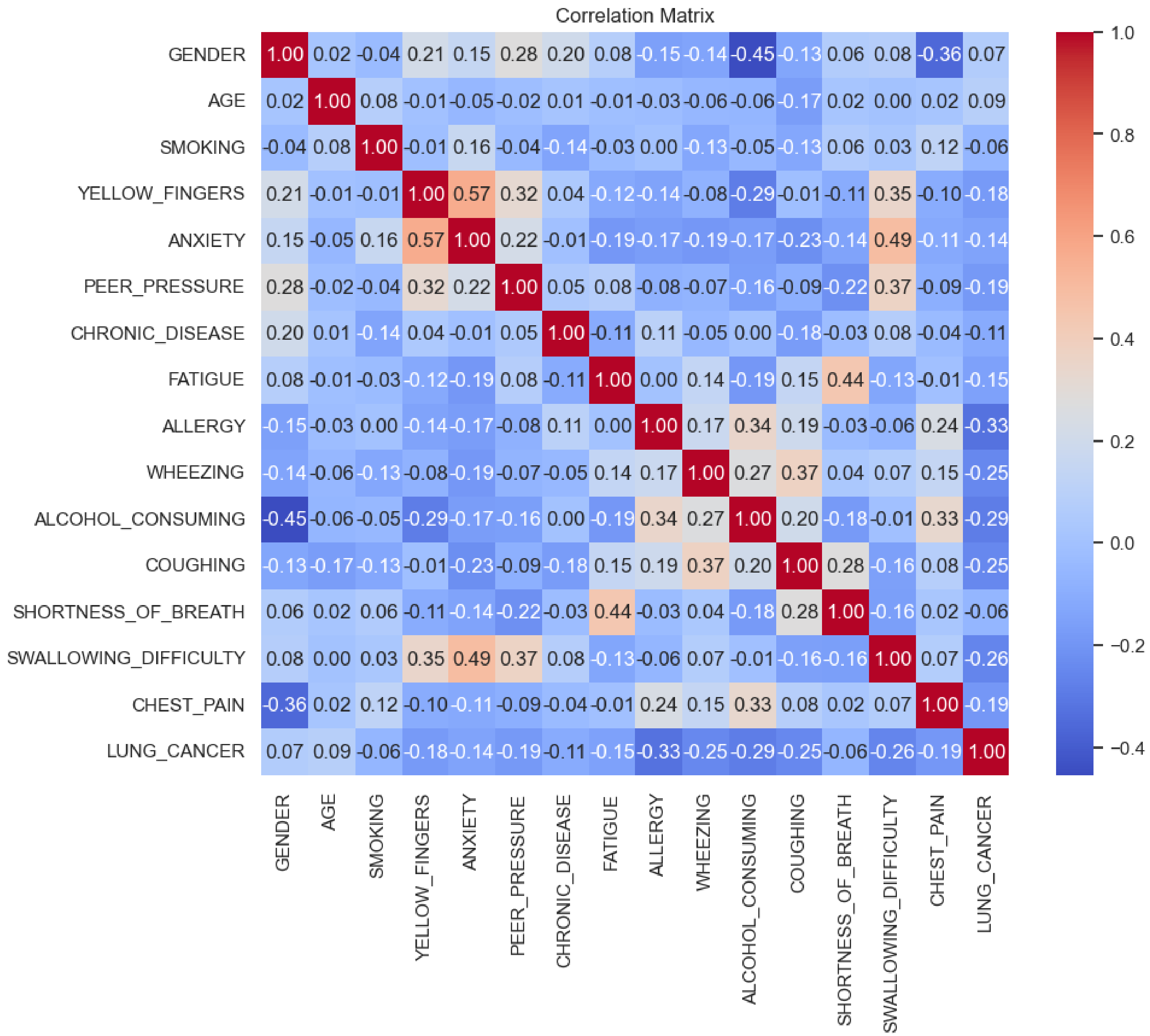
- The dataset is not balanced in terms of age, with a majority of older individuals.
- A minority of individuals actually have lung cancer.
- Anxiety is strongly correlated with yellow fingers, and moderately correlated with swallowing difficulty. Shortness of breath is moderately correlated with fatigue.
- Age, allergy, alcohol consumption, and peer pressure are among the most important features.

Exploratory Data Analysis



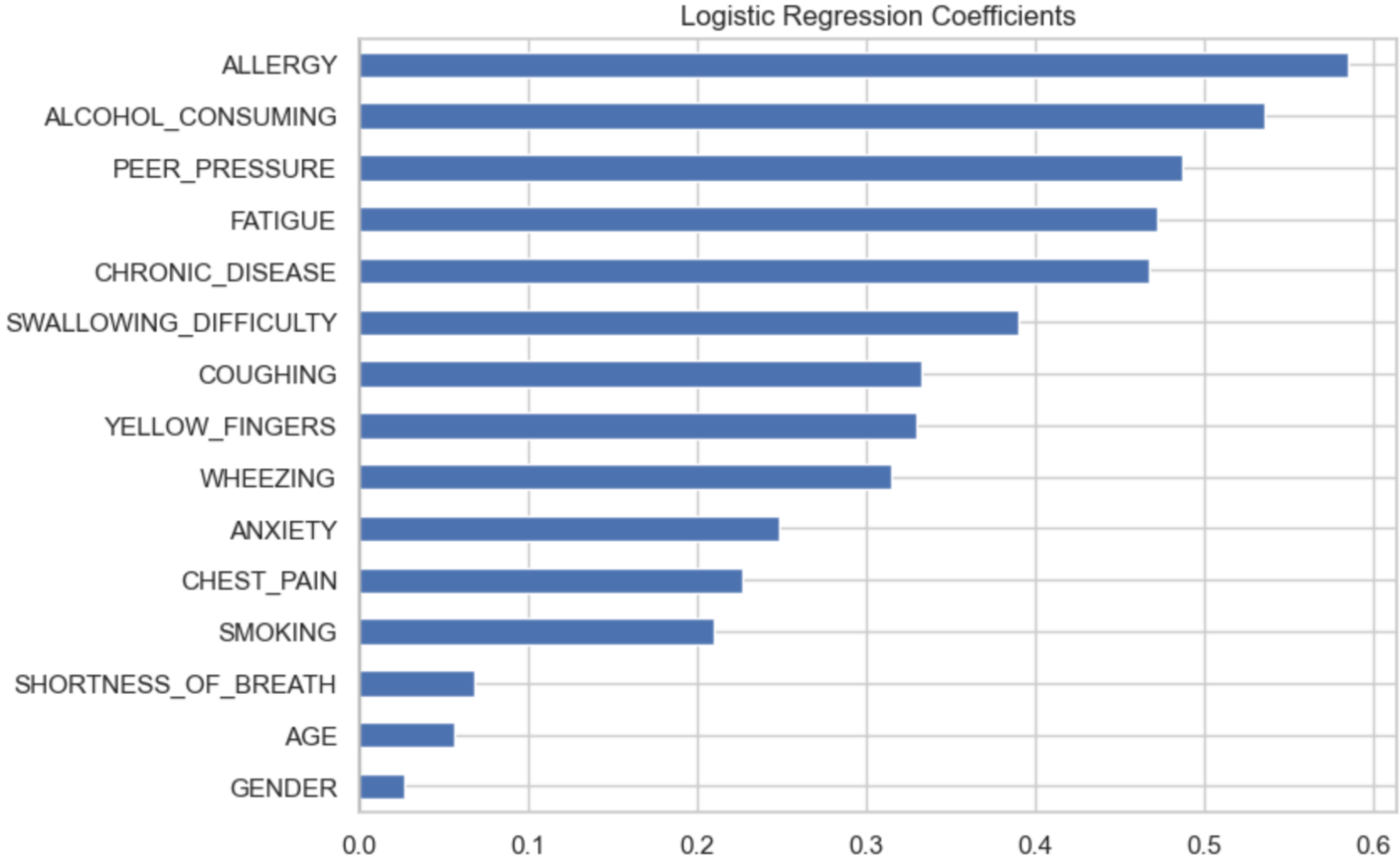

Target Class Balance
Count
250
200
150
100
50
0
No
Yes
LUNG_CANCER


GENDER
SMOKING
YELLOW_FINGERS
ANXIETY
PEER_PRESSURE
CHRONIC_DISEASE
FATIGUE
ALLERGY
WHEEZING
ALCOHOL_CONSUMING
COUGHING
SHORTNESS_OF_BREATH
SWALLOWING_DIFFICULTY
CHEST_PAIN
count
No
Yes
0
1

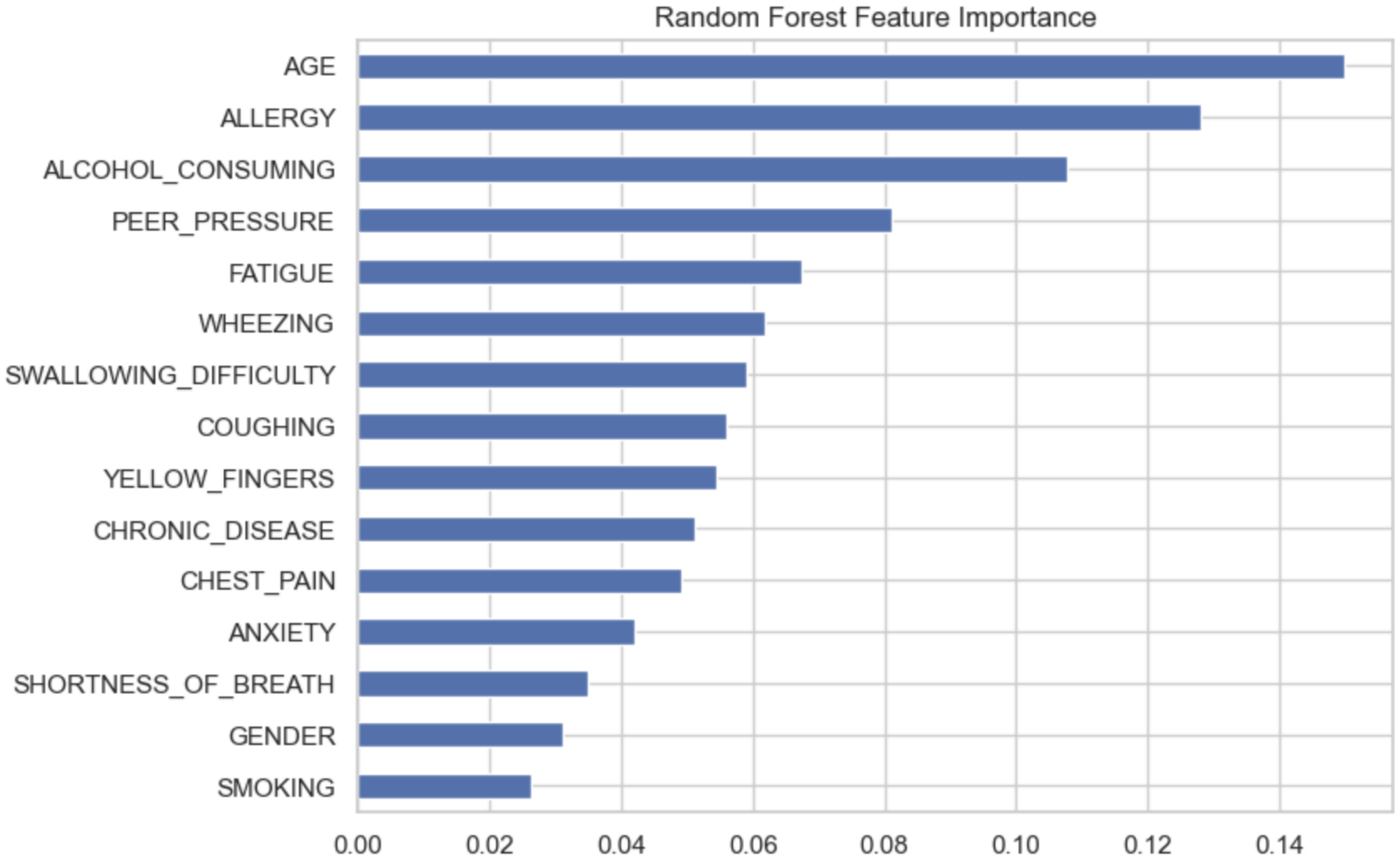


Correlation Matrix
GENDER 1.00 0.02 -0.04 0.21 0.15 0.28 0.20 0.08 -0.15 -0.14 -0.45 -0.13 0.06 0.08 -0.36 0.07
AGE 0.02 1.00 0.08 -0.01 -0.05 -0.02 0.01 -0.01 -0.03 -0.06 -0.06 -0.17 0.02 0.00 0.02 0.09
SMOKING -0.04 0.08 1.00 -0.01 0.16 -0.04 -0.14 -0.03 0.00 -0.13 -0.05 -0.13 0.06 0.03 0.12 -0.06
YELLOW_FINGERS 0.21 -0.01 -0.01 1.00 0.57 0.32 0.04 -0.12 -0.14 -0.08 -0.29 -0.01 -0.11 0.35 -0.10 -0.18
ANXIETY 0.15 -0.05 0.16 0.57 1.00 0.22 -0.01 -0.19 -0.17 -0.19 -0.17 -0.23 -0.14 0.49 -0.11 -0.14
PEER_PRESSURE 0.28 -0.02 -0.04 0.32 0.22 1.00 0.05 0.08 -0.08 -0.07 -0.16 -0.09 -0.22 0.37 -0.09 -0.19
CHRONIC_DISEASE 0.20 0.01 -0.14 0.04 -0.01 0.05 1.00 -0.11 0.11 -0.05 0.00 -0.18 -0.03 0.08 -0.04 -0.11
FATIGUE 0.08 -0.01 -0.03 -0.12 -0.19 0.08 -0.11 1.00 0.00 0.14 -0.19 0.15 0.44 -0.13 -0.01 -0.15
ALLERGY -0.15 -0.03 0.00 -0.14 -0.17 -0.08 0.11 0.00 1.00 0.17 0.34 0.19 -0.03 -0.06 0.24 -0.33
WHEEZING -0.14 -0.06 -0.13 -0.08 -0.19 -0.07 -0.05 0.14 0.17 1.00 0.27 0.37 0.04 0.07 0.15 -0.25
ALCOHOL_CONSUMING -0.45 -0.06 -0.05 -0.29 -0.17 -0.16 0.00 -0.19 0.34 0.27 1.00 0.20 -0.18 -0.01 0.33 -0.29
COUGHING -0.13 -0.17 -0.13 -0.01 -0.23 -0.09 -0.18 0.15 0.19 0.37 0.20 1.00 0.28 -0.16 0.08 -0.25
SHORTNESS_OF_BREATH 0.06 0.02 0.06 -0.11 -0.14 -0.22 -0.03 0.44 -0.03 0.04 -0.18 0.28 1.00 -0.16 0.02 -0.06
SWALLOWING_DIFFICULTY 0.08 0.00 0.03 0.35 0.49 0.37 0.08 -0.13 -0.06 0.07 -0.01 -0.16 -0.16 1.00 0.07 -0.26
CHEST_PAIN -0.36 0.02 0.12 -0.10 -0.11 -0.09 -0.04 -0.01 0.24 0.15 0.33 0.08 0.02 0.07 1.00 -0.19
LUNG_CANCER 0.07 0.09 -0.06 -0.18 -0.14 -0.19 -0.11 -0.15 -0.33 -0.25 -0.29 -0.25 -0.06 -0.26 -0.19 1.00
GENDER AGE SMOKING YELLOW_FINGERS ANXIETY PEER_PRESSURE CHRONIC_DISEASE FATIGUE ALLERGY WHEEZING ALCOHOL_CONSUMING COUGHING SHORTNESS_OF_BREATH SWALLOWING_DIFFICULTY CHEST_PAIN LUNG_CANCER
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

Logistic Regression Feature Importance



Random Forest Feature Importance



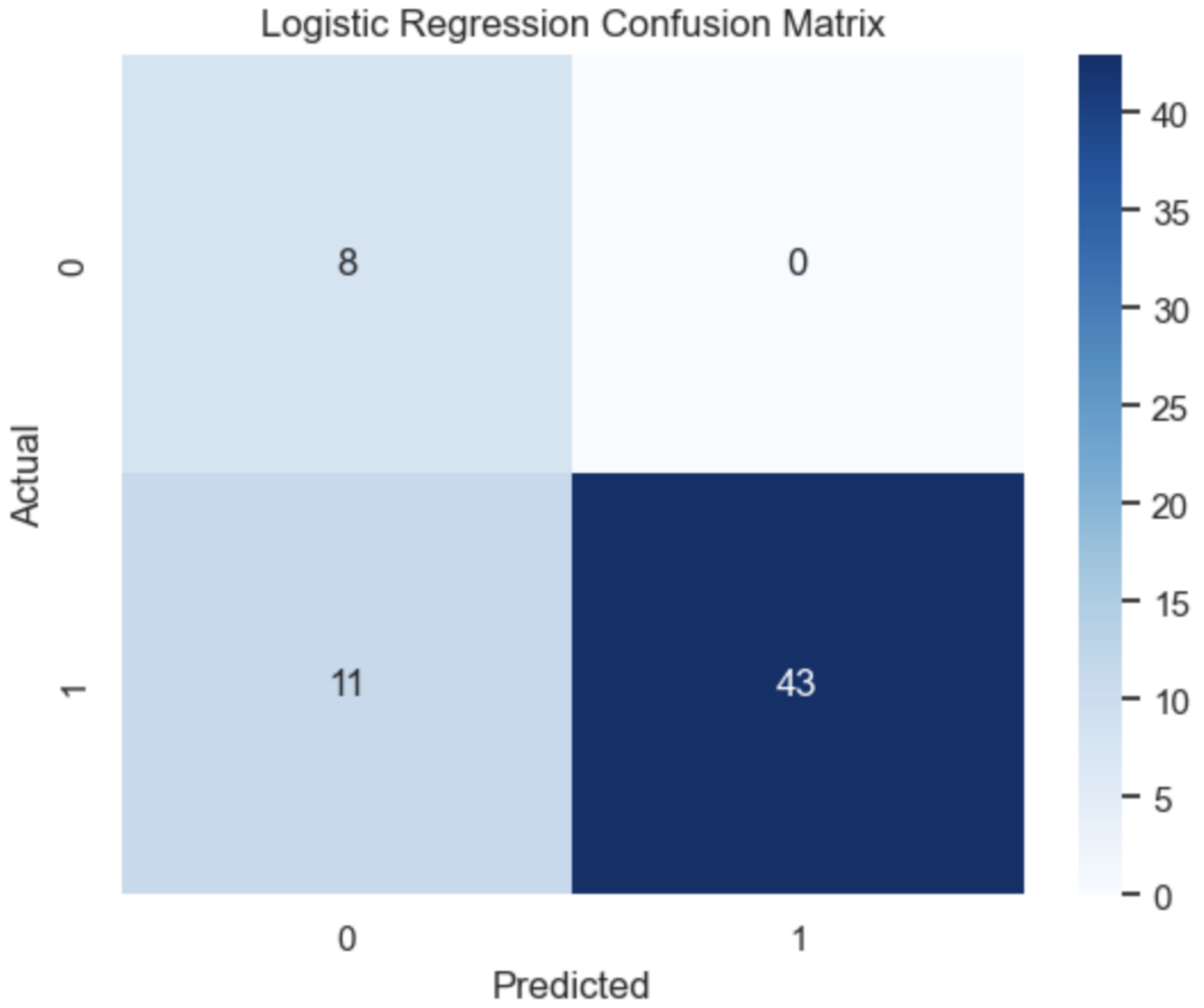
Modeling

- Logistic regression and random forest models are utilized.
- Hyperparameter tuning is performed, and the best parameters are included (e.g. clf_c = 0.001).
- L2 regularization is applied by default through logistic regression, helping to prevent overfitting by penalizing large coefficient values.
- Feature engineering is performed by generating interaction terms through polynomial feature expansion, capturing nonlinear relationships between variables.

Results

- Logistic regression performed without hyperparameter tuning, regularization and feature engineering has an ROC AUC that is 0.01 lower than that of the optimized model.
- Optimized LR misclassifies ~18% of all samples with an accuracy of 82.3%.
- Random forest has an AUC comparable to that of non-optimized LR but with an accuracy of almost 89%.
- This means that random forest ranks predictions roughly the same as LR, but makes more correct predictions at the given threshold.

Logistic Regression Confusion Matrix

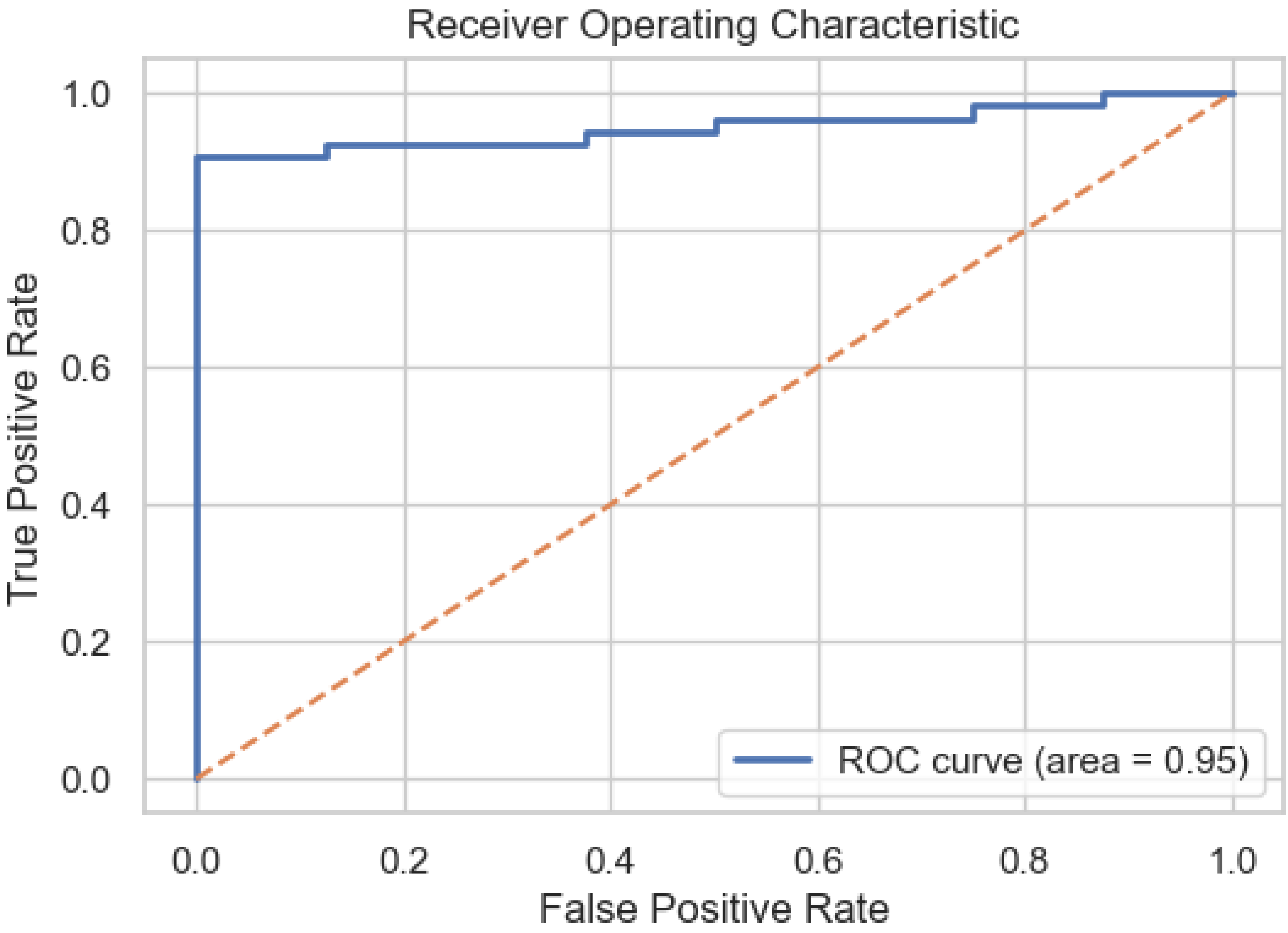


Logistic Regression Analytics

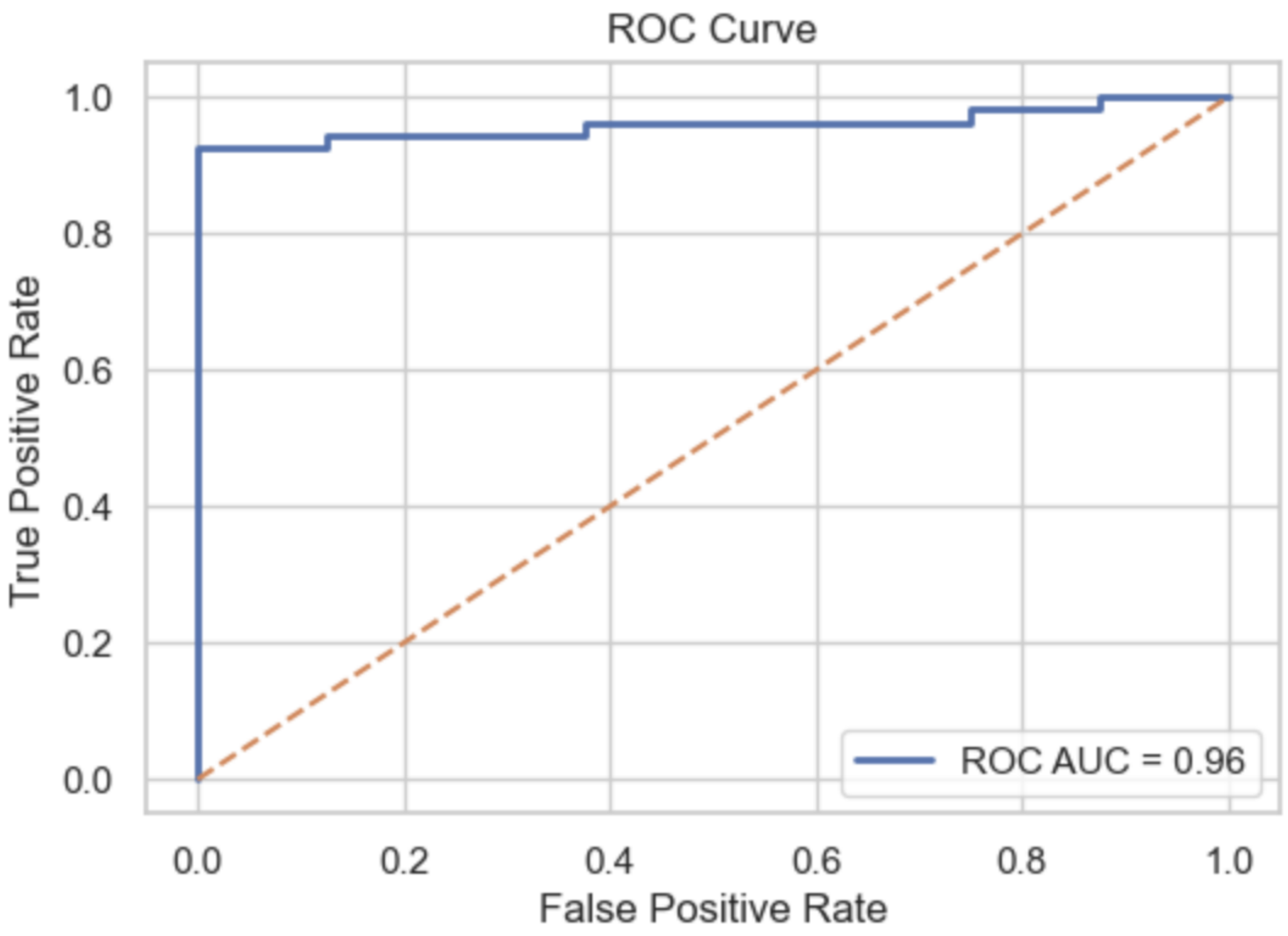
```
Classification Report
              precision    recall  f1-score   support

           0       0.42      1.00      0.59         8
           1       1.00      0.80      0.89        54

    accuracy                           0.82        62
   macro avg       0.71      0.90      0.74        62
weighted avg       0.93      0.82      0.85        62

Accuracy: 0.823
```

```
Best params: {'clf__C': 0.001, 'clf__l1_ratio': 0.0, 'clf__penalty': 'l2', 'poly__interaction_only': True}
Best CV AUC: 0.931
```

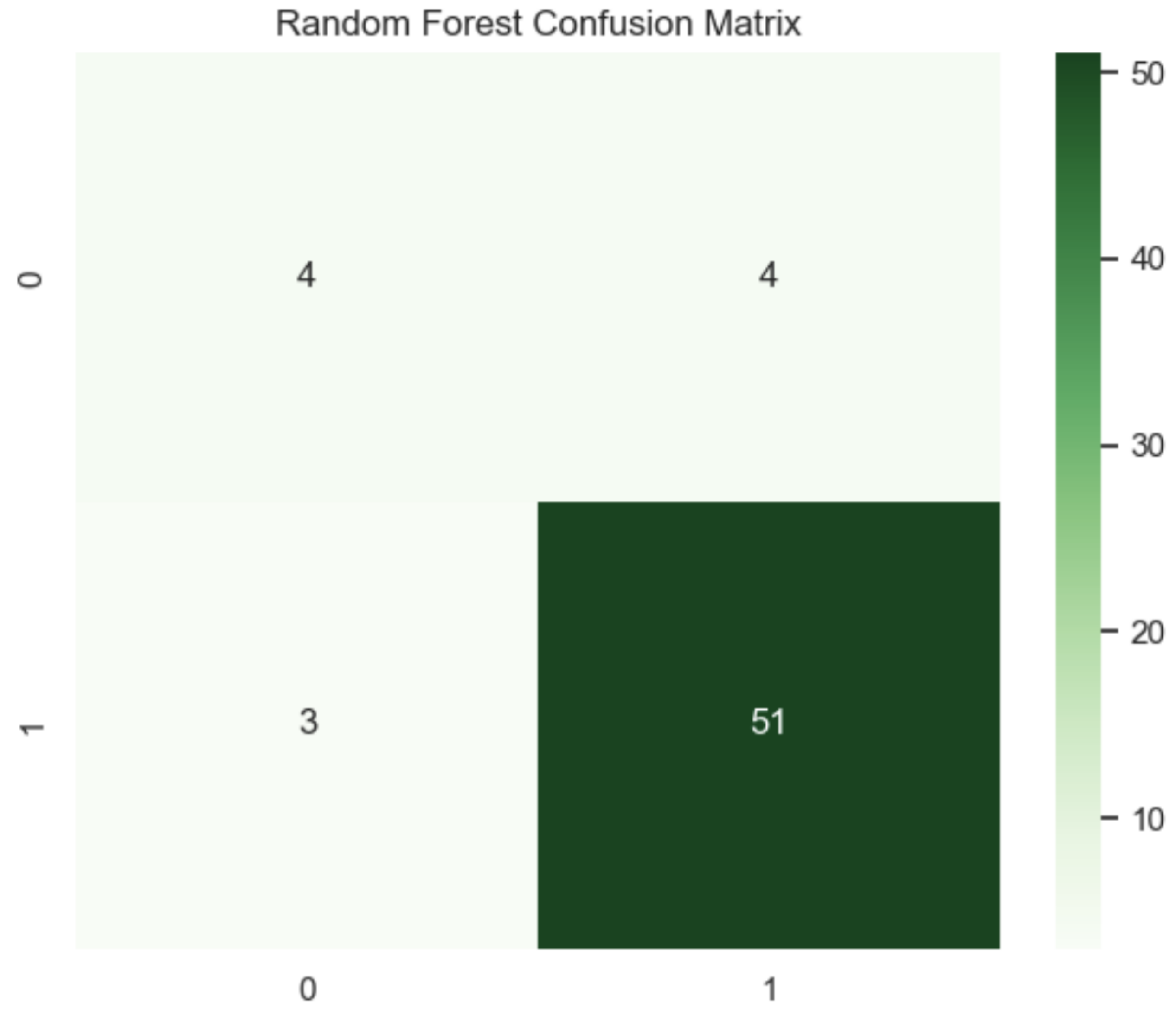
LR Non-Optimized ROC



LR Optimized ROC



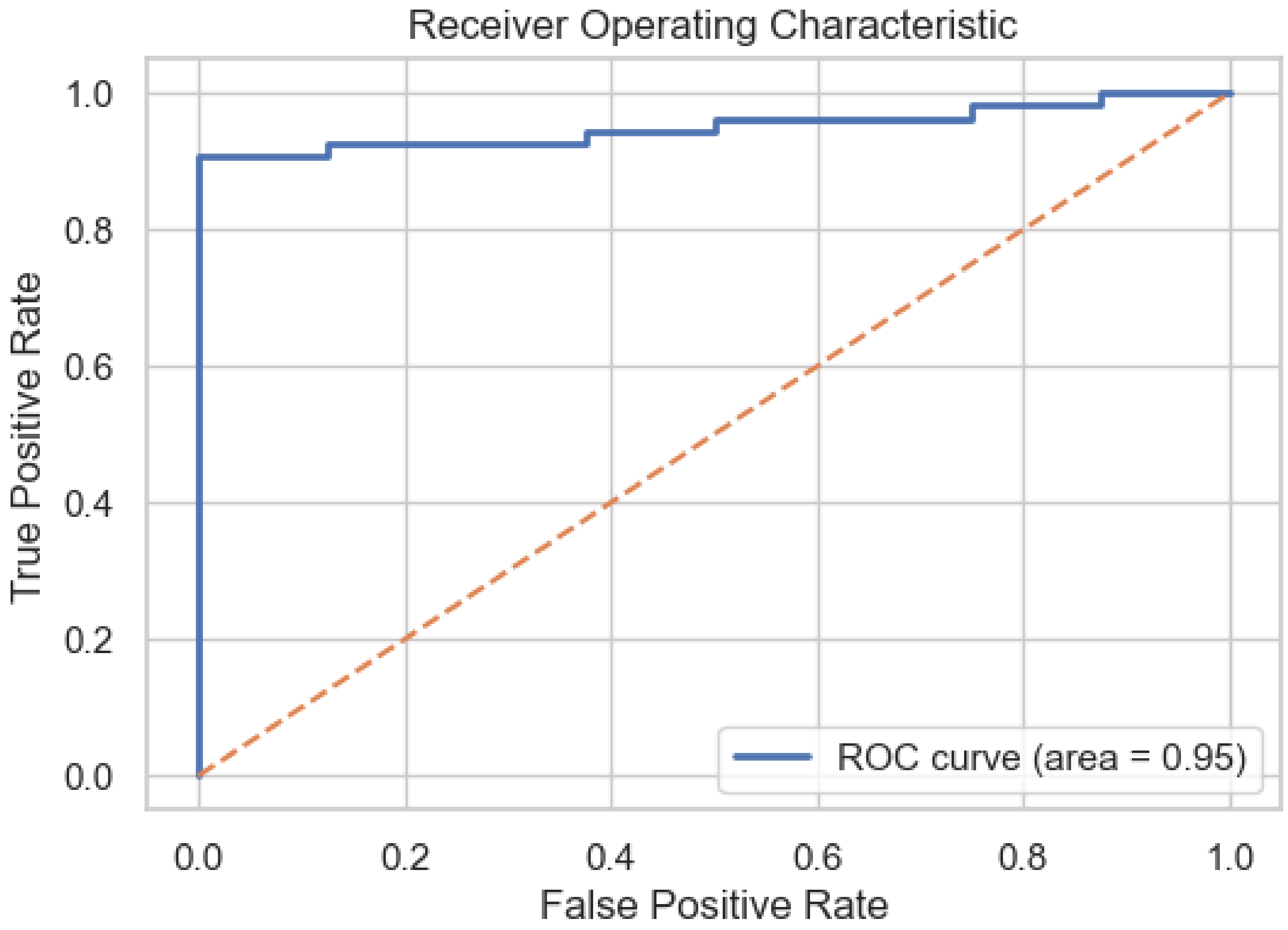
Random Forest Confusion Matrix



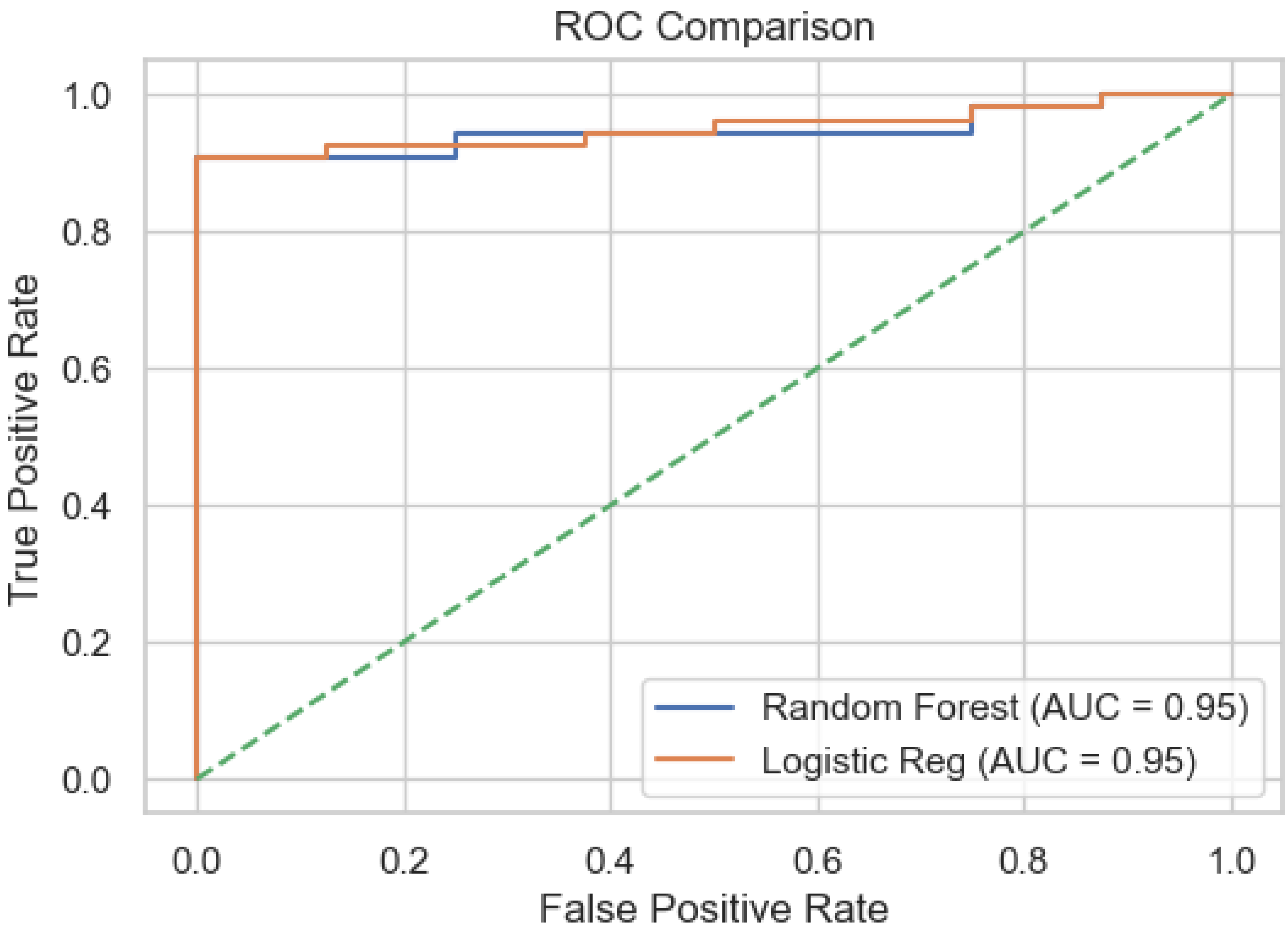
Random Forest Analytics

```
Classification Report (Random Forest)
              precision    recall  f1-score   support

           0       0.57      0.50      0.53         8
           1       0.93      0.94      0.94        54

    accuracy                           0.89        62
   macro avg       0.75      0.72      0.73        62
weighted avg       0.88      0.89      0.88        62

Accuracy: 0.887
```

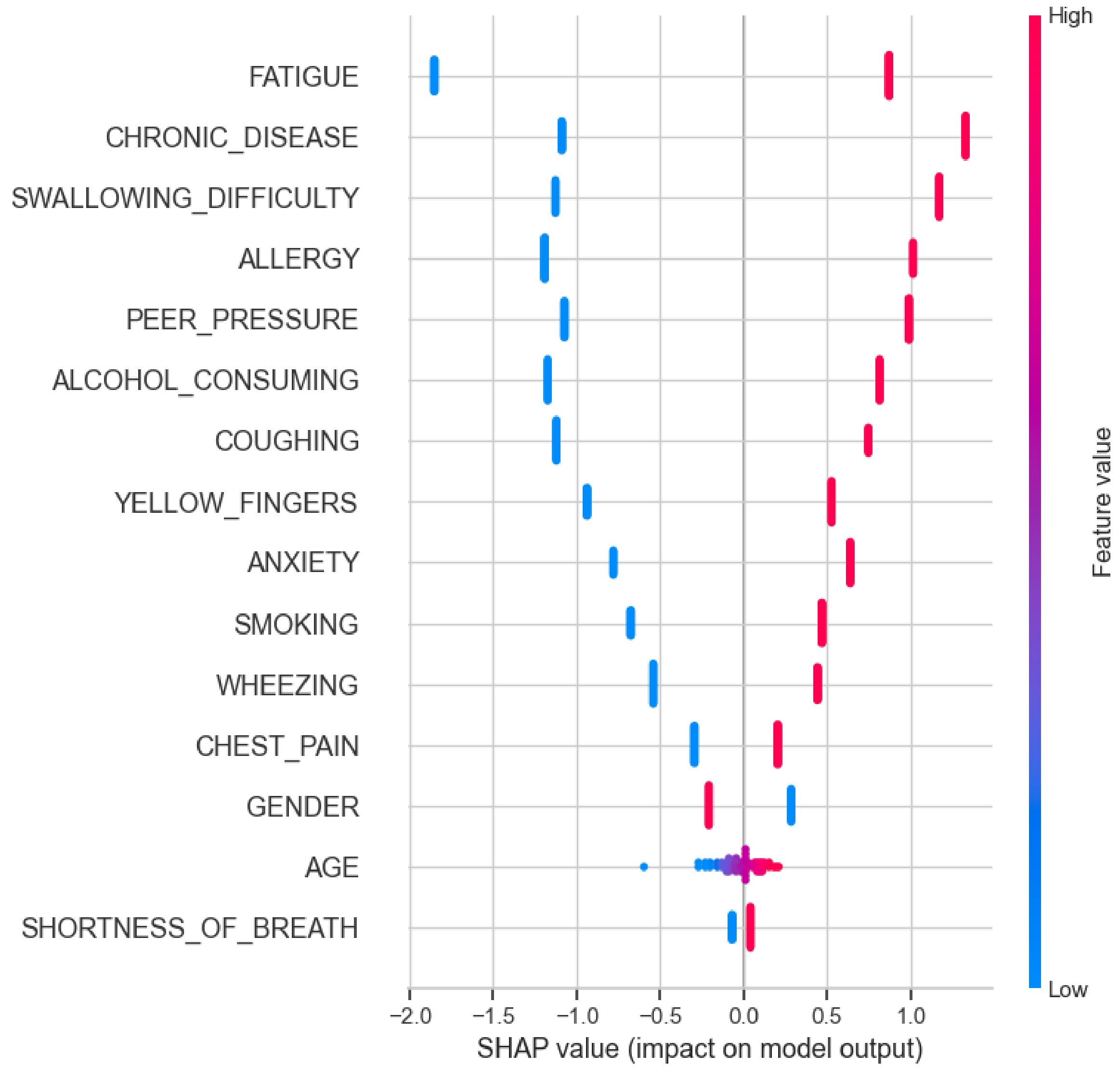
Random Forest Non-Optimized ROC



Model Comparison



Logistic Regression SHAP Summary



Random Forest SHAP Summary

Conclusion

- Logistic regression with tuned regularization parameters reach relatively high AUC and accuracy, while Random Forest slightly improves AUC but at the cost of interpretability.
- The highest-weighted features include SMOKING, AGE, and respiratory symptoms such as COUGHING, SHORTNESS_OF_BREATH, and CHEST_PAIN.
- This aligns with the medical literature, stating that tobacco exposure is the leading risk factor for lung cancer and respiratory symptoms are common in patients that have a diagnosis.

Limitations

- The dataset size (n ~ 300) is relatively small; collecting more samples would improve the accuracy of these models.
- Survey responses might be self-reported, which can add bias and potentially mis-annotate the data.
- Other AI/ML techniques should be explored (e.g. Gradient Boosting, XGBoost) along with more involved hyper-parameter optimization with nested CV.
- A web application should be developed to allow clinicians to input survey responses and receive risk scores; this will address the data gap outlined previously.

Q&A

Thank You!