Introduction

Machine Learning (ML) has taken its place in the world impacting decision making, routine tasks, and professional insights across all industries. Although most of ML's influence has been on information technology and forecasting, people continue to advance ML into dependable resources in other fields. In healthcare and risk management, ML is starting to develop and impact the industries enabling actuaries to become more efficient with evaluating uncertainty.

Mortality modeling can achieve considerably high levels with ML's hand in predictive analysis. Actuaries depend on traditional approaches such as broad demographic classification and historical claims experience, however, modern ML methods allow actuaries to integrate deeper clinical and behavioral information. This could lead the industry to more advanced risk segmentation and quicker identification of insurance holders with high risk. As costs continue to increase in healthcare, trends have become less predictable and thus increasing the need for tools that improve mortality assessment.

This paper develops on how well ML models can predict mortality outcomes for insurance policy holders using a heart failure dataset. The heart failure Dataset being utilized in this research includes 368 patient observations and 60 predictors collected at the Faisalabad Institute of Cardiology. Some variables in the dataset are Clinical history, blood test results, cardiovascular markers, follow-up frequency, and demographic data. Mortality is the binary outcome variable with factors 0 for survival and 1 for death. Heart disease has had a major impact on mortality and insurance costs allowing it to be a relevant dataset for developing a strong model. By evaluating both model performance and the variables most associated with mortality, this study highlights how emerging analytical techniques can complement traditional

actuarial judgment. The goal of this research is to identify which learning model most accurately predicts patient mortality and to determine which variables are associated with the outcome.

Statistical Background and terminology

Traditionally, actuaries have relied on life tables, generalized linear models, and credibility methods for accurate estimates to develop insurance rates. The idea is not to replace these methods, however, to add more to these methods in order to further the actuarial process.

During exploratory data analysis (EDA) we had to tackle problems such as data cleaning and multicollinearity.

Before discovering models, we began Exploratory Data Analysis (EDA) on the data by cleaning the data. This consisted of removing or replacing missing observations in the dataset and restructuring the data to a tidy form. Some of the methods for handling missing data consisted of removing the row or replacing the missing value with the average value of the column. Following data cleaning, we removed unnecessary variables. One method we removed unnecessary variables was by evaluating multicollinearity with the Variance Inflation Factor (VIF). Multicollinearity is caused by highly correlated variables leading to redundant or unnecessary information. It is important to evaluate multicollinearity because of its ability to maintain powerful results while minimizing complexity in the model.

The models we used in this research were Logistic Regression, Stochastic Gradient Descent Support Vector Machine (SGD SVM), and random forest. Logistic Regression is a supervised ML model utilized to predict binary variables. Since Logistic Regression is a fundamental model, we utilized it as a baseline model in order to have an understanding of what minimum we need to achieve with the more advanced models. Stochastic Gradient Descent Support Vector Machine (SGD SVM). The SGD part of the model is the iterative optimizer to

minimize the loss function. In other words, it adjusts the weights of the parameters in the model in order to maximize the model's efficiency. The SVM part is the main model itself with SGD supporting it. This model was important in the research because it helped us identify how significant numerical predictors are. The random forest ML model is important because of its robust performance with classification and regression. The foundation of the random forest model is building many decision trees and then combining them for better decision making and accuracy.

We used various metrics and visualizations to evaluate the models. Accuracy was our main metric for analysis and it was supported by confusion matrix, f1-score, recall, precision, and the ROC curve. The confusion matrix gives us the number of correct answers, wrong answers, and what types they are. F1-score allows us to analyze the model performance in respect to class imbalance. Recall is the true positive rate in the confusion matrix. Precision measures how often the model is right when predicting the positive outcome. The ROC curve visualizes the true positive rate against the false positive rate in order to give us a better picture of the model performance.

Methodology

When going through this research, we utilized R and Python in order to align with researchers' skillsets. During the EDA process, R and Rstudio was used. For visualization, we utilized the ggplot2 package. For the models, we utilized Python. The imported libraries for the Python work were pandas, scikit-learn, NumPy, seaborn, matplotlib.pyplot, and statsmodels.

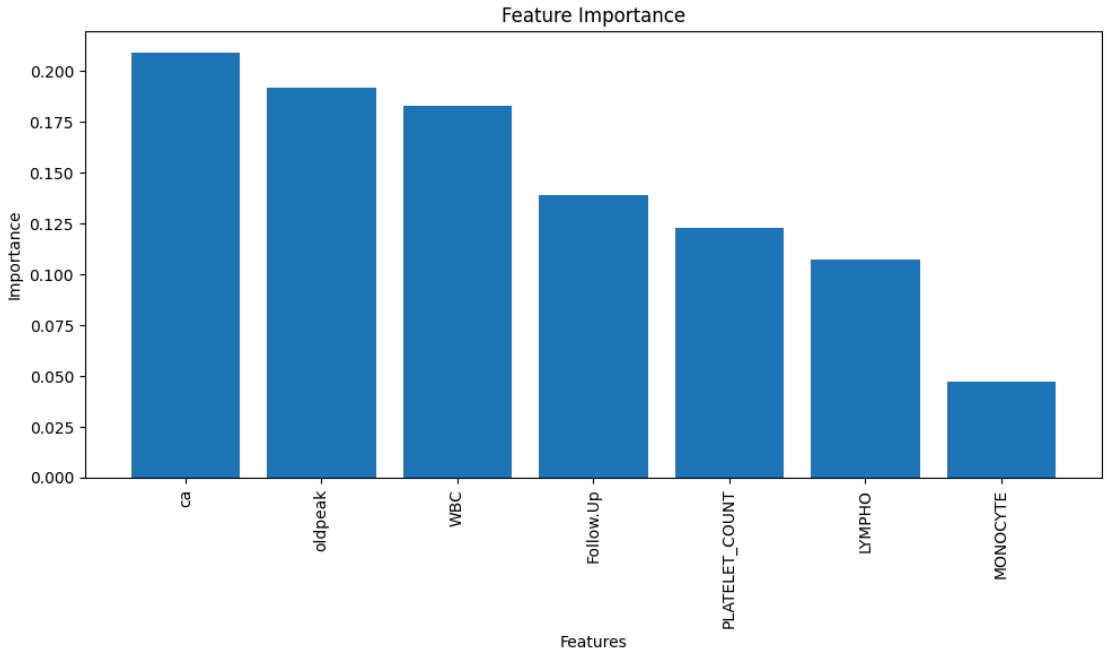
After analyzing the VIFs and removing most of the variables, research on some variables were conducted in order to understand the safe and healthy ranges allowing us to transform the variables. After the EDA process, we moved forward with the following 7 predictor variables:

White Blood Cell Count (WBC), Platelet Count, Number of Major Vessels, Number of follow-ups, Lymphocyte Ratio, Monocyte Percentage, and St-Segment Depression.

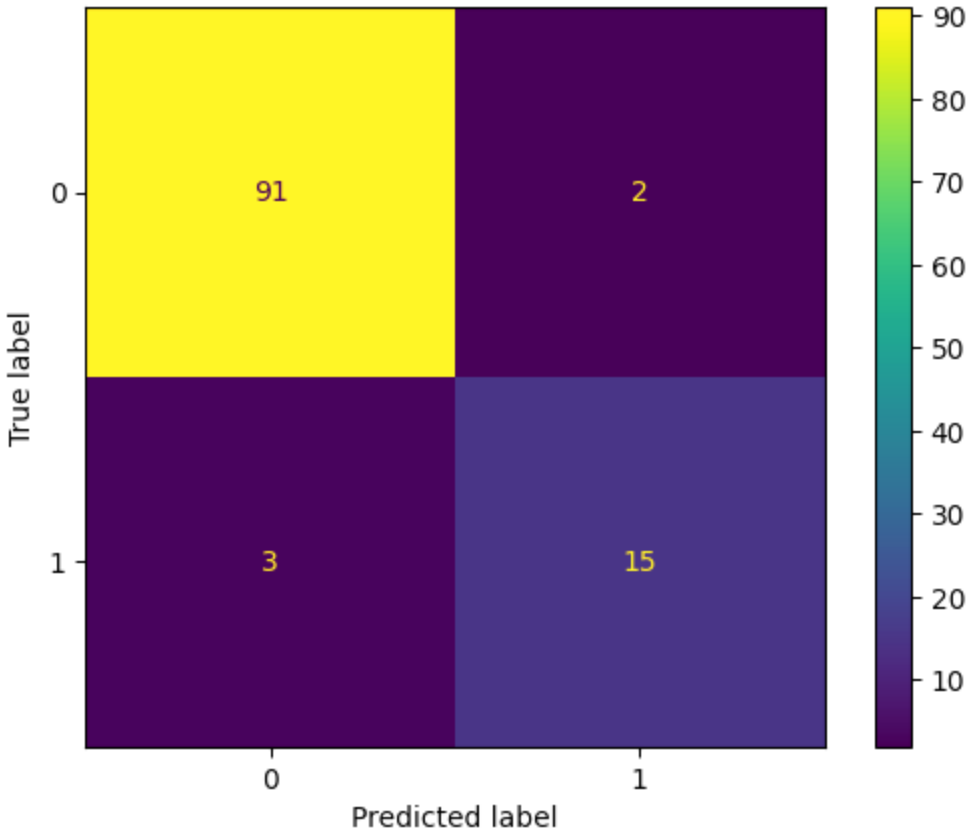
When prepping the data for models, we split the data 70/30 for training and testing in order to efficiently run the models. Due to the data being small, we set regularization for the logistic regression model to be .5 telling the model to trust the training set less and focus more on prioritizing a more generalized model. We assigned .001 to the learning rate for the SGD SVM model allowing the model to more incremental improvements rather than quicker unstable changes. For random forests, we performed hyperparameter tuning in order to maximize the model efficiency and accuracy. We utilized GridSearchCV and found the optimal tree count for accuracy to be n = 10. This method also kept us from leading to extensive computational time.

Results

Using the feature importance plot below, we learned about the relationship between the predictors and the response variable. Lower levels of WBC were found to be associated with a higher survival rate. A higher Lymphocyte ratio was found to be related with a higher mortality risk. The number of major vessels was significant for high-risk patients. A patient can be living without all of their major vessels, however, having all major vessels is significant towards survival rate. The final relationship found was the follow-up frequency with mortality risk, however, this may be because patients living long will have more follow up appointments.



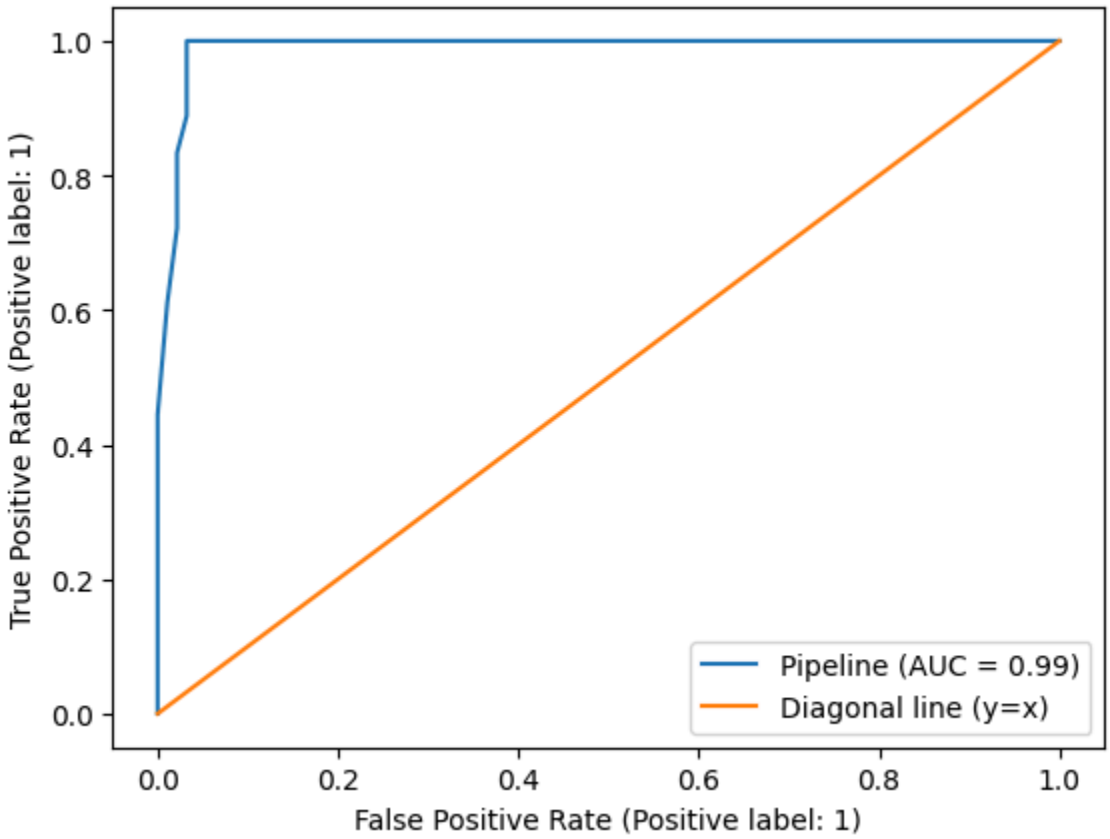
Results showed random forest leading with 98.2% accuracy. Following Random Forest was Logistic Regression and SGD SVM with 82.9% and 82.0% accuracy. A shortcoming for SGD SVM was its inability to efficiently handle smaller datasets. After identifying the random forest model as the best model, we continued to analyze the model by observing the confusion matrix and metrics.



		precision	recall	f1-score
Mortality 0 = Died 1 = Alive	0	1.00	0.98	0.99
	1	0.90	1.00	0.95
accuracy				0.98
macro avg		0.95	0.99	0.97
weighted avg		0.98	0.98	0.98

The confusion matrix correctly predicted 106 out of 111 observations. The model correctly predicted 91 cases for death and 15 cases for survival. with only five total misclassified cases. This shows that the model was able to figure out the difference between the two mortality outcomes while having balanced predictability for both classes. Following the confusion matrix we reviewed the metrics to identify other characteristics in the model. The metrics for the model support the confusion matrix and from the f1-score, we see a balance between precision and

recall for both binary outcomes in the predictor variable. After looking at the confusion matrix and the metrics, we completed the analysis by reviewing the ROC-curve.



The ROC-curve shows strong performance for differentiation in the response variable classes. The .99 AUC means the model is almost perfect with differentiating death with survival. The curve in the plot allows us to conclude that the model is significantly better than random guessing.

Conclusion

This study evaluated how the ability of modern machine learning models can improve mortality prediction when compared with more traditional methods used by actuaries. The random forest model showed the best performance, accurately predicting the outcome 98.2% of the time and holding an ROC AUC of .99. For an actuarial audience, this means that mortality

depends on some combination of the factors instead of a simple linear relationship. In practice, this could revolutionize the identification process of high-risk policyholders with better precision when provided the proper data.

The analysis also shed light on multiple variables that were impactful in prediction, specifically the seven variables used in the predictive modeling. These results from the research agree with the broader principle that the response variable is impacted by inflammation and cardiovascular condition. Although traditional actuarial methods are still essential to the process, machine learning methods could transform the industry by improving claim forecasting and health risk segmentation. A shortcoming of the research was the small dataset encouraging future researchers to follow up with more data and analysis before applying these methods to real actuarial processes.

Appendix

Link to data, code, and outputs:

https://drive.google.com/drive/folders/1_cEVywDsWwyoQP-edyV1kDEAThkHiPid?usp=drive_link

Predictor	Description
White Blood Cell Count (WBC)	measures leukocytes in the blood to detect infection, inflammation, or bone marrow disorders
Platelet Count	measures the number of platelets (thrombocytes) in your blood to evaluate clotting ability
Number of Major Vessels	The blood vessels that are connected to the heart
Number of follow-ups	The number of following visits with the doctor after the initial
Lymphocyte Ratio	the percentage of lymphocytes relative to the total white blood cell count
Monocyte Percentage	measures the proportion of monocytes among total white blood cells (typically 2-8%),

	acting as an inflammatory biomarker
St-Segment Depression	typically signifies subendocardial ischemia (inadequate oxygen to the heart muscle) or ventricular hypertrophy, often presenting as horizontal or downsloping depression in contiguous leads