AUGUST 2023

NAVIGATING THE NEXUS OF HEALTH AND WEALTH

A Survival Analysis Model for Chronic Disease Management, Cost Efficiency and Quality Care in Health Insurance

Authors: Scott Howard, Mu Miao, Esther Xu

TABLE OF CONTENT

01. PROJECT INTRODUCTION

02. DATA SOURCES

03. ANALYSIS & FINDINGS

04. BUSINESS IMPLICATIONS & CONCLUSIONS

01.

PROJECT INTRODUCTION

A SOLUTION FOR HEALTH INSURERS

- Health insurance organizations have a dual purpose: to ensure the health and well-being of their participating members while maintaining financial sustainability and profitability.
- Cox proportional hazards regression (Cox-PH): identify members diagnosed with a chronic disease and most at risk of death.
- Key areas of opportunity: Enhanced Preventive Care and Improved Quality of Care, Profitability and Resource Optimization, Financial Forecasting and Improved Risk Stratification, and Federal Reimbursements.

MEDICAL CLAIMS AND ADMINISTRATIVE DATA

- Health insurance companies can utilize claims and administrative data systems to gain valuable insights into their members' health trends, behaviors, and needs.
- Claims data: includes information on medical diagnoses, treatments, and costs and provides a detailed view of a patient's medical history.
- Administrative data: enrollment, demographics, and billing. This data can be used to understand the socio-economic factors influencing health, streamline billing processes, and enhance customer service.

A FRAMEWORK FOR ANALYZING TIME-TO-EVENT DATA

Proportional Hazards Assumption

Cox Proportional Hazards Regression (Cox-PH)

Model Interpretation

Extensions

- PyCox
- DeepSurv

02.

DATA SOURCES

CMS DATA ENTREPRENEURS' SYNTHETIC DATA

The Data Entrepreneurs' Synthetic Public Use Files (PUF) was developed with the intention of offering a genuine set of public claims data while ensuring the protection of Medicare beneficiaries' confidential health information.

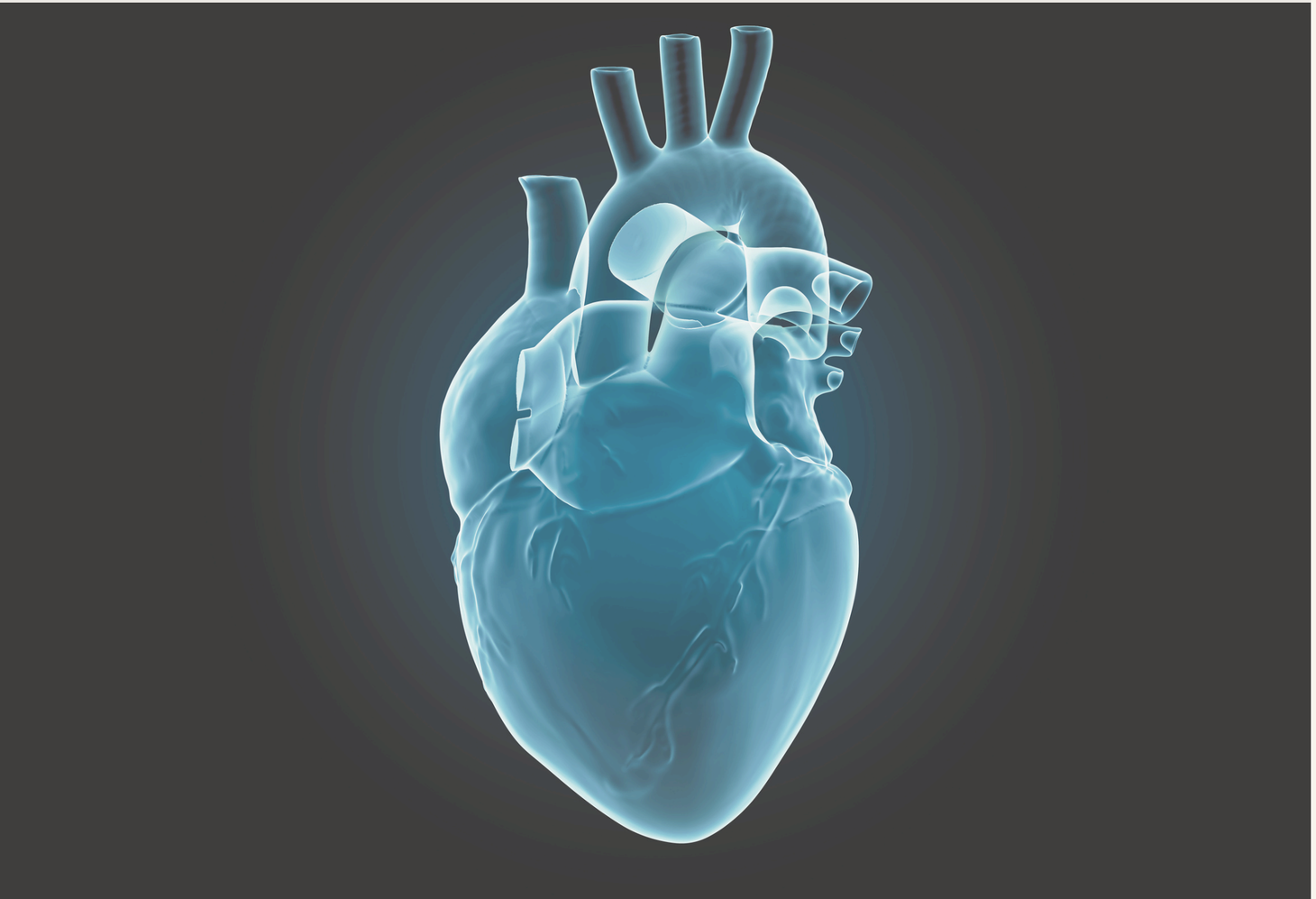
Five Types of Datasets

1. Beneficiary Summary Files
2. Inpatient Claims
3. Outpatient Claims
4. Carrier Claims
5. Prescription Drug Events

KEY FIELDS

- Death
- Age at death (or age at censored period)
- Sex
- Race
- Hospitalizations
- Total medications (as well as total grams dispensed)
- Total outpatient visits (as well as copay costs)

THE WORCESTER HEART ATTACK STUDY (WHAS) DATA



The Worcester Heart Attack Study data is an extensive dataset collected as part of a population-based surveillance study focusing on acute myocardial infarction (AMI) in residents of the Worcester metropolitan area in Massachusetts. The dataset spans three decades from 1975 to 2005. The data provides a comprehensive and in-depth look at the incidence rates, demographic and clinical characteristics, treatment practices, and outcomes related to AMI over a 30-year span.

KEY FIELDS

- Death
- Length of follow-up (observation period)
- Age
- BMI
- Sex
- Heart Rate
- Blood Pressure

03.

ANALYSIS & FINDINGS

CMS Dataset

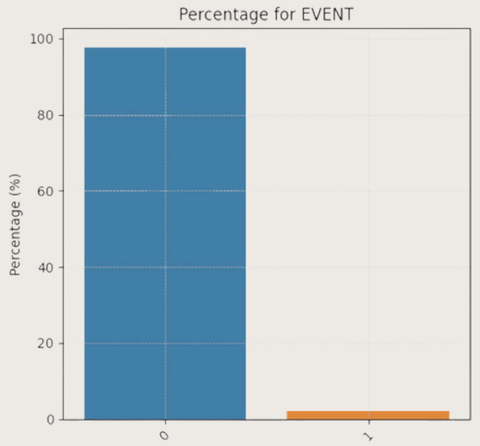
DATA SPLIT TECHNIQUE

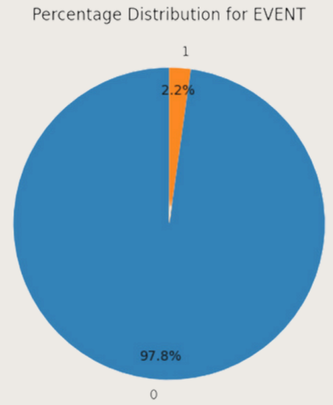
The "deaths" event in our data is significantly imbalanced. To address this, we implemented corrective measures across both the training and test data sets.

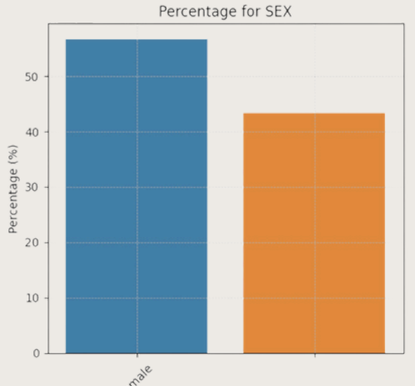
Solutions Implemented:

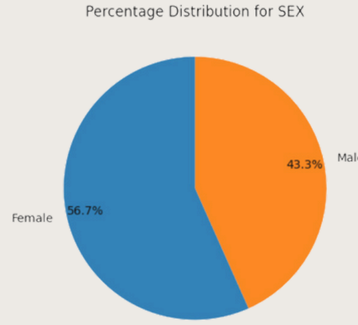
1. **Class Weights Technique:**
   Balance the impact of each class during model training

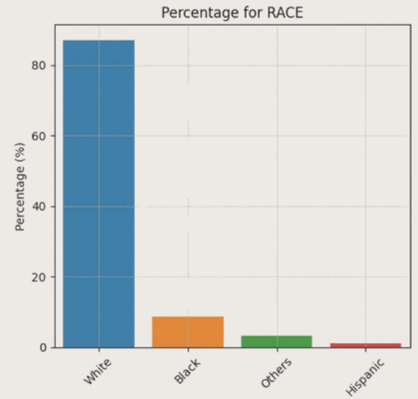
2. **Stratification:**
   Ensures balanced representation of *EVENT, RACE, and SEX.*

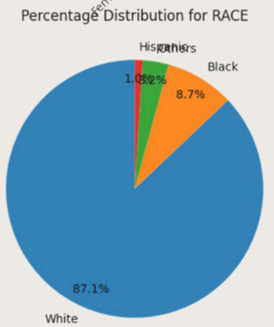










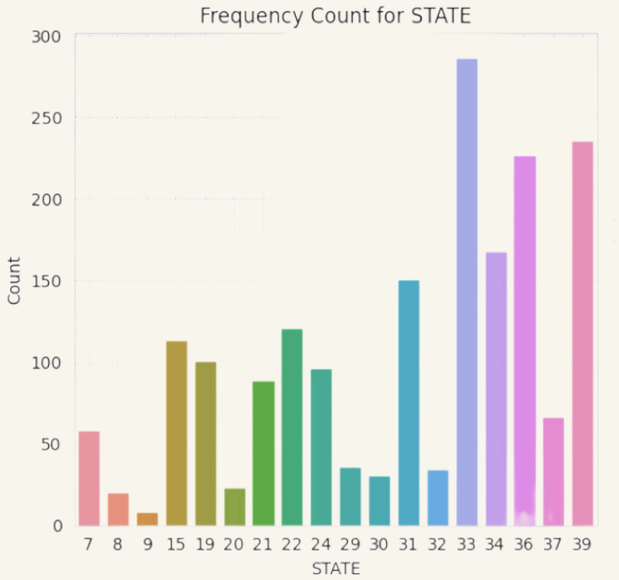


Univariate Analysis

DATA EXPLORATION

Geographical Variance:

- States exhibit different prevalence rates of heart disease.

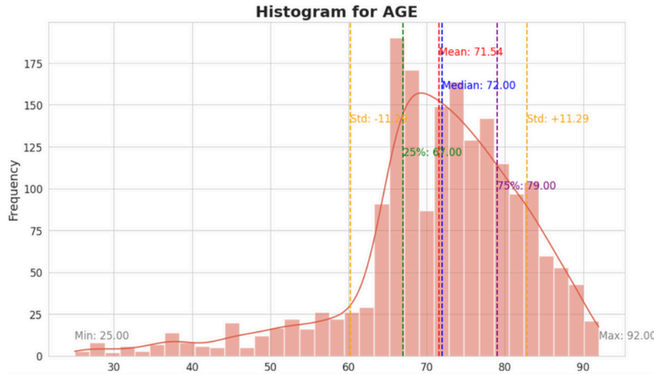


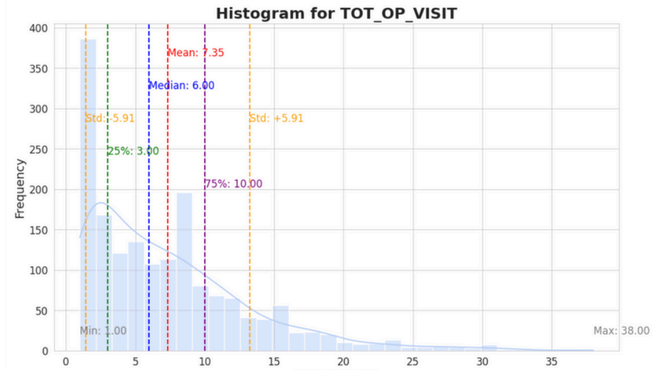
Age Factor:

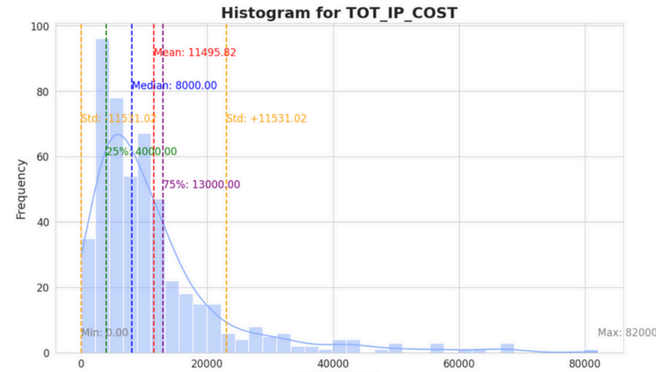
- Heart disease is notably more prevalent in individuals aged 60 and above。

Distribution of 'tot_op_visit' and 'cost' Variables

- Both variables demonstrate a right-skewed distribution.
- Most observations are clustered on the left, with fewer high-value observations on the right.





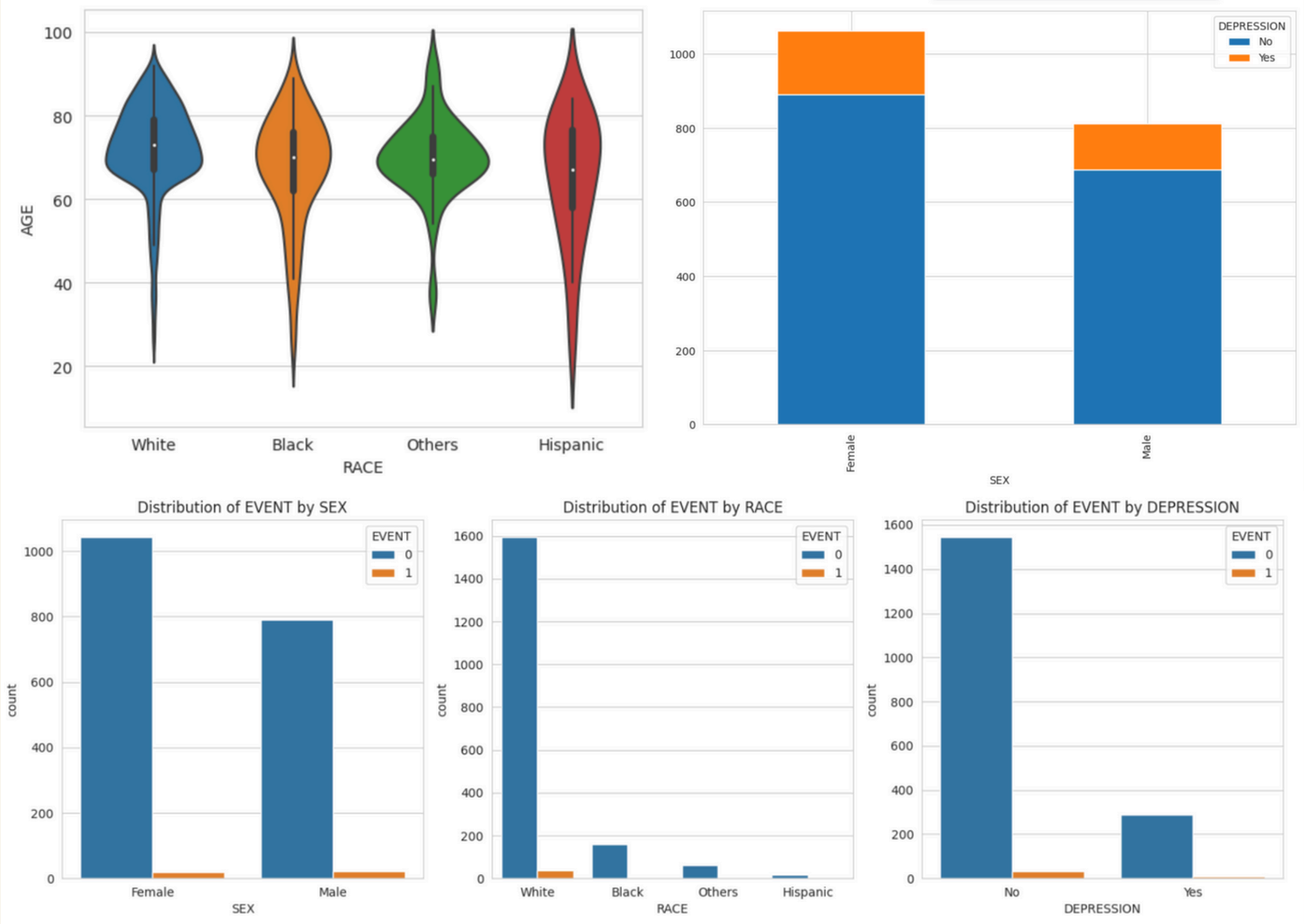


Bivariate Analysis

DATA EXPLORATION

Heart Disease Rates Across Races:

- Median Prevalence:
  - Almost uniform across all races.
- Observation for Hispanics and Blacks:
  - Distinctively longer tails in distribution.
  - Indicates the onset of heart disease at a younger age compared to other races.
  - Notably, death rates are very low for these groups.



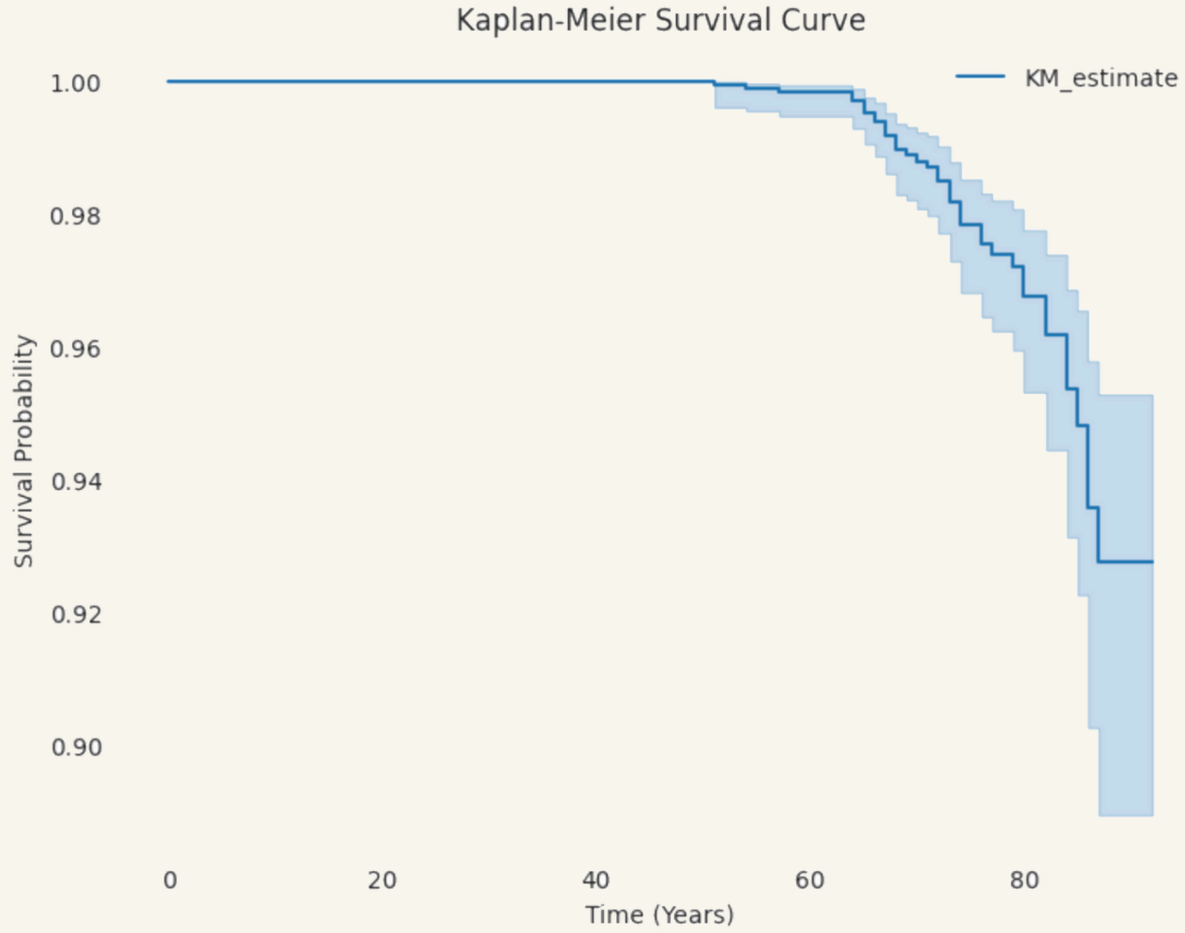
Gender Disparities in Heart Disease and Depression:

- Depression Prevalence:
  - Significantly higher in men compared to women.
- Heart Disease and Depression
  - Chart indicates that death is not determined by the presence of depression.
  - Preliminary inference: Male heart disease patients have a higher mortality rate than their female counterparts.

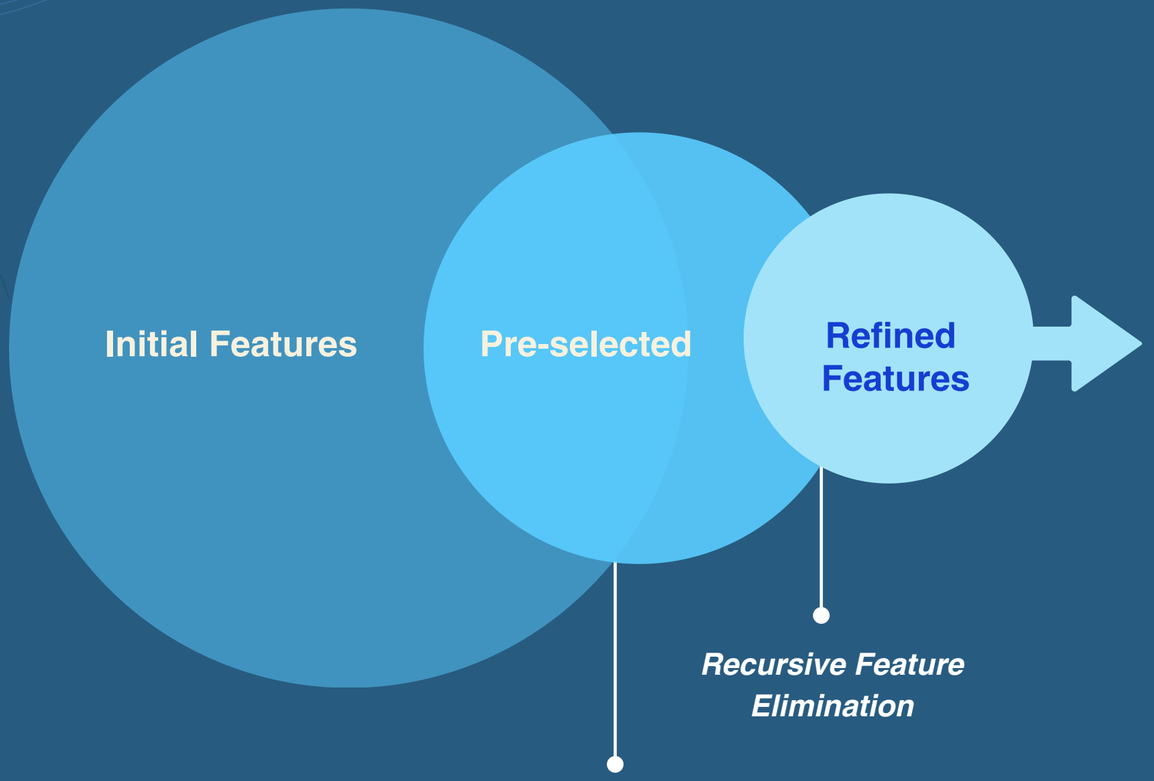
Survival Analysis

DATA EXPLORATION

Kaplan-Meier Survival Curve plots survival probabilities over time for different age groups. Expect a steeper decline for those aged 60+ compared to younger age groups, indicating a higher risk of death.



DATA PRE-PROCESSING



- *Replaced missing TOT columns‘ values with 0*
- *Convert 2 value categories into binary variables*
- *Convert race into dummy variables*

SEX

Depression

TOT_*

- TOT_IP_ADMSN
- TOT_OP_VISIT
- TOT_MEDS

Race_*

- RACE_HISPANIC
- RACE_OTHERS
- RACE_WHITE

COX-PH MODEL

Model Training

Concordance Index (c-index) on training dataset: ***0.9662***

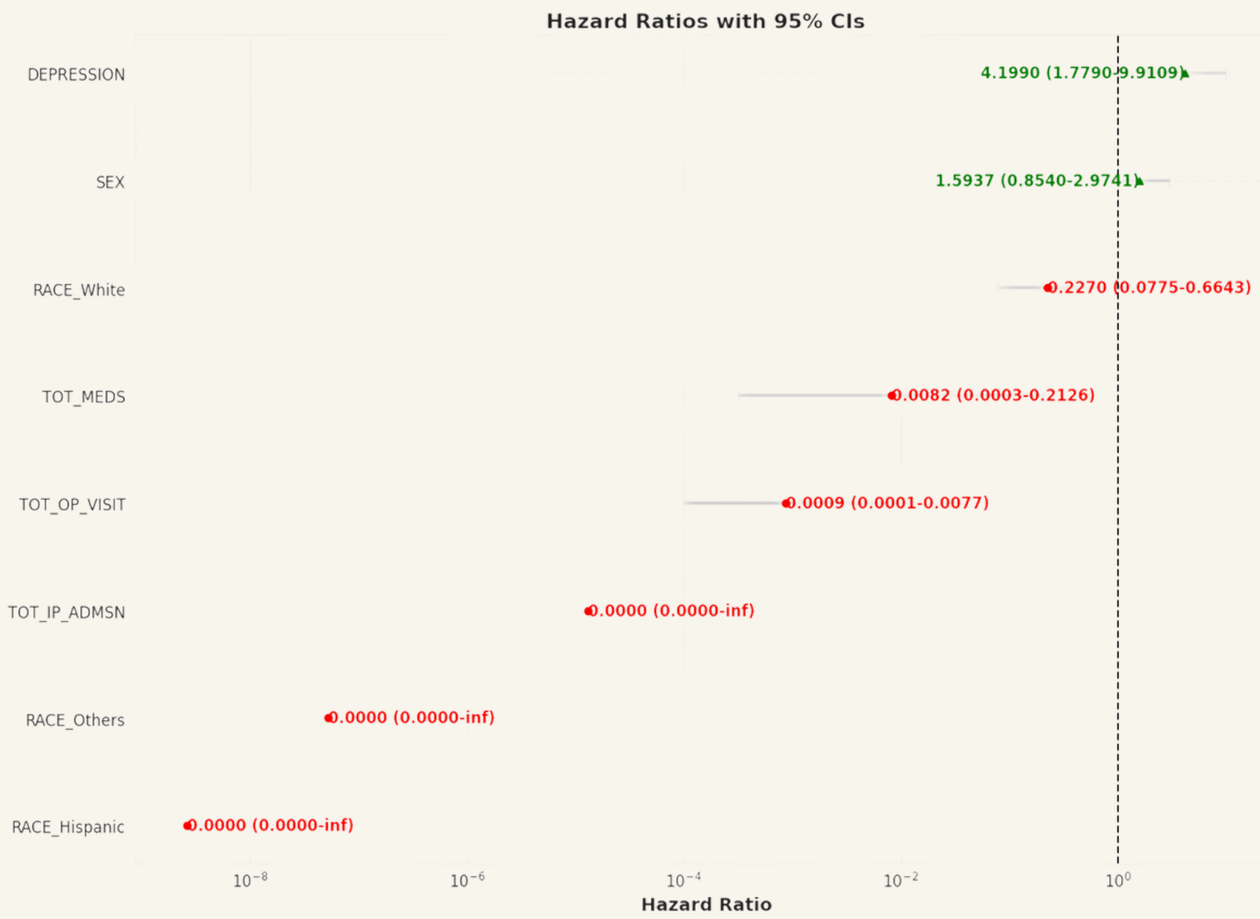
Model Validation

- **K-folds Cross Validation:** Average C-index across 5 folds: ***0.9617***
- **Bootstrapping with Cox-PH:**95.0 confidence interval of C-index: ***0.96 - 0.97***

Model Evaluation

Concordance Index (c-index) on test dataset: ***0.9092***

COX-PH MODEL INTERPRETATION



Key Takeaways from the Graph:

SEX (1.5937):
- Men have 59.37% higher risk of death from heart disease.
- Confidence Interval: 0.85340 to 2.9741.

DEPRESSION (4.1990):
- 4.2 times higher risk of death for depressed patients.
- Confidence Interval: 1.7790 to 9.9109.

TOT_IP_ADMSN (0.0000129):
- The risk of death decreases with each inpatient admission.
- Confidence Interval: 0 to infinity (interpret with caution).

TOT_OP_VISIT (0.0009):
- Each outpatient visit significantly reduces risk.
- The importance of regular check-ups is highlighted.

TOT_MEDS (0.0082):
- More medications reduce death risk.
- Indicates better care or treatment compliance.

RACE_Hispanic (0.00000000259) & RACE_Others (0.0000000522):
- Near-zero risk, but confidence goes to infinity.
- Indicates potential data scarcity or category underrepresentation.

RACE_White (0.2270):
- 77.3% risk reduction compared to reference race.
- Calls for deeper dive into care access or social determinants.

Business Implications - Patient Care:
- Prioritize interventions for high-risk groups: Men & those with depression.
- Advocate for regular outpatient visits.
- Ensure suitable medication regimes to decrease mortality.
- This provides a succinct representation of the data and its implications. Adjust the layout and design of the slide to ensure visual clarity and impact.

PYCOX MODEL
Implementation and Results

Configuration:
- Input Features: Based on dataset dimensions.
- Network Architecture:
- Layers: Input -> 64 -> 64 -> Output.
- Activation: ReLU.
- Regularization: Dropout (30%).
- Purpose: To model non-linearity and learn complex patterns.

Details:
- Optimizer: Adam.
- Batch Size: 256.
- Epochs: 20.

Steps:
- Convert tensors back to numpy for evaluation.
- Compute baseline hazards using a trained model.
- Predict survival curves for test data.
- Evaluate the model using C-index.

Key Result:
- C-index: 0.6725.
- Interpretation: Ability of the model to rank survival times.

Conclusion:
- Successfully implemented Cox's model with neural enhancements.
- The model's performance was evaluated using C-index.

DEEPSURV MODEL
Implementation and Results

Purpose:

- Implement DeepSurv using PyTorch.
- Convert data frames to tensors, define model architecture, and train the model.
- Evaluate model performance using the Concordance Index (C-index).

Key Components:

- Data Preparation:
  - Convert data frames into PyTorch tensors: Features, Time, and Events.
- Model Architecture:
  - Multi-layer Neural Network with dropout and ReLU activations
  - Output predicts hazard ratios.
  - Custom loss function: CoxPHLoss.
- Training:
  - Adam optimizer, learning rate decay, gradient clipping.
  - Early stopping based on validation loss.
- Results:
  - A model trained for 500 epochs.
  - Achieved a C-index of 0.3675 on test data.

03.

ANALYSIS & FINDINGS

Worcester Heart Attack Dataset

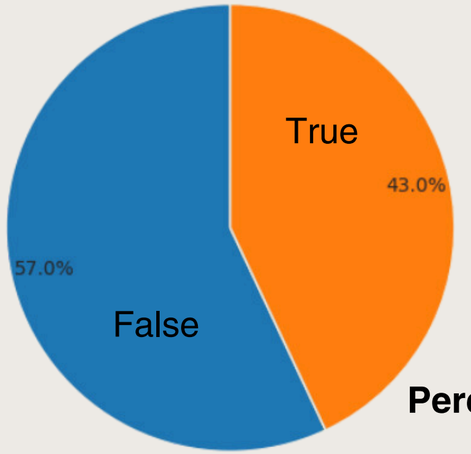
DATA SPLIT TECHNIQUE

Ensure balanced representation of gender and age group in training and test datasets.

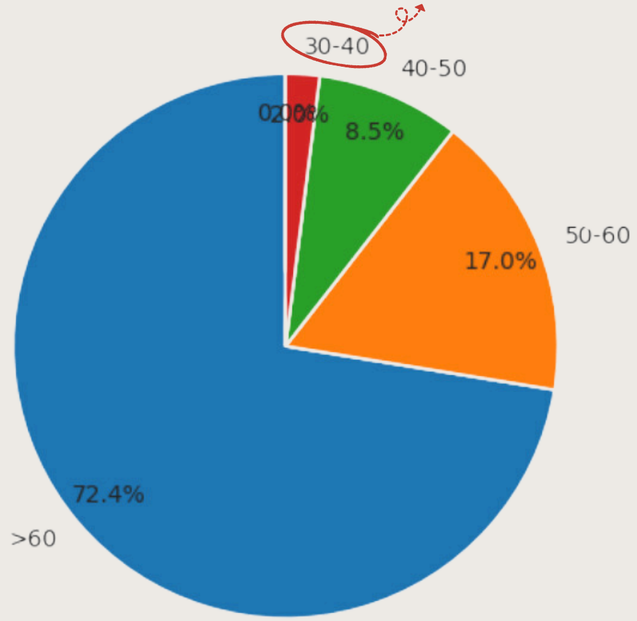
Solutions Implemented:

- **Approach:**
  - Employed StratifiedShuffleSplit to maintain proportionate representation.
  - This technique requires at least 2 instances for each class to work effectively.
- **Challenge:**
  - The gender-age combination '30-40' appeared only once in the dataset.
  - Solution: Removed this particular entry to allow stratified splitting.

Percentage Distribution for fstat



Percentage Distribution for age group



Univariate Analysis

DATA EXPLORATION

Age by Genders:

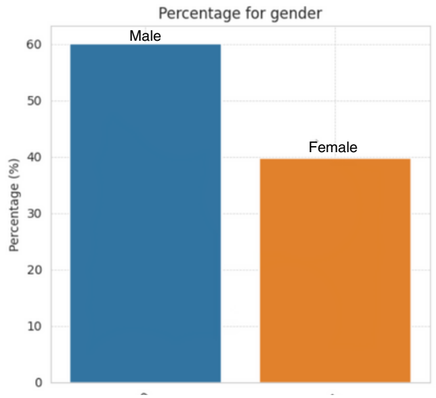
- 60% Male
- 40% Female

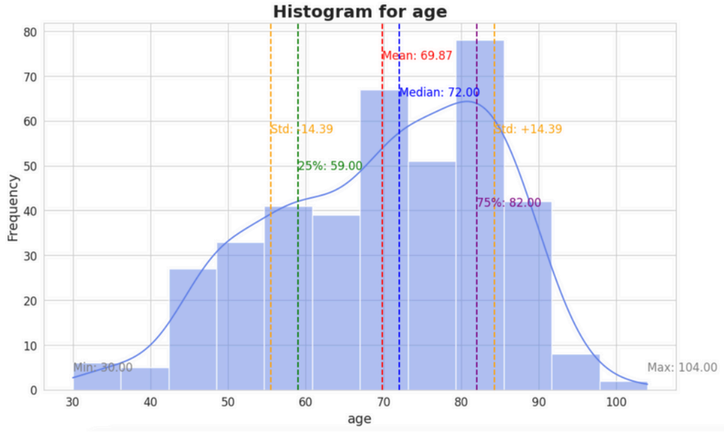
Age Distribution:

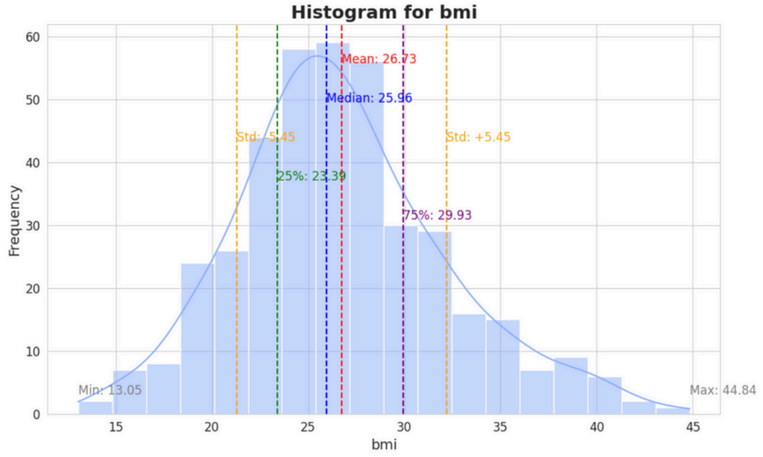
- Senior group focused
- The average and median are near 70

Body Mass Index Distribution

- BMI centers within a health range
- The overweight group outsizes the underweight group





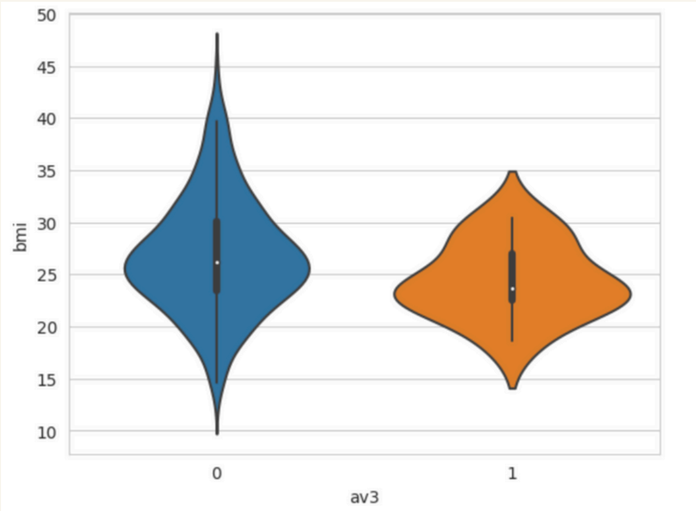


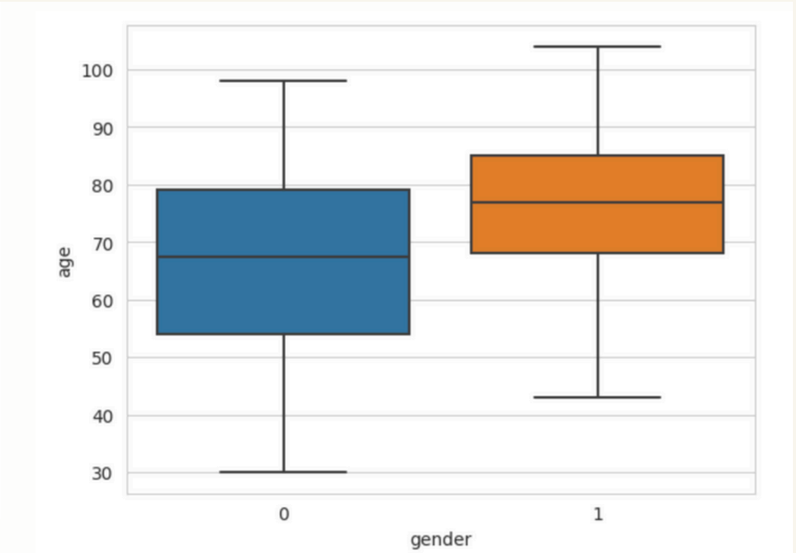
Bivariate Analysis

DATA EXPLORATION

Third-degree atrioventricular(av3):

- BMI:
  - Almost uniform w/o av3.
- Event Count:
  - Small sample size in the av3 positive group
  - False events are larger in the av3 negative group



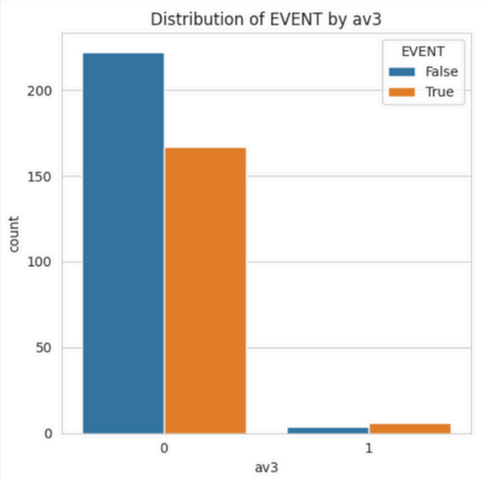


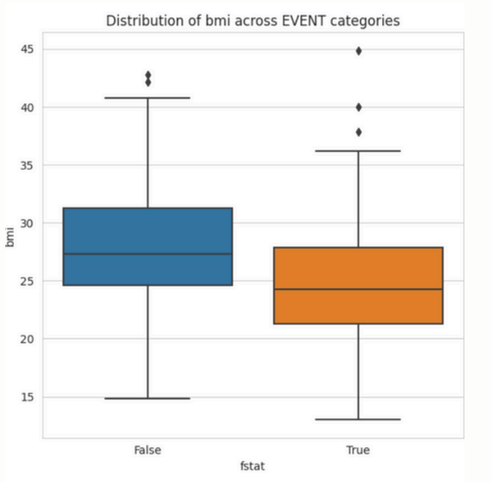
Genders vs. Ages:

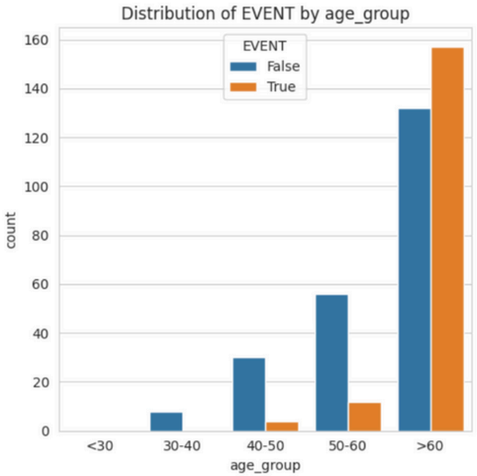
- Samples of females are older than males

BMI impacts on death:

- BMI near 25 shows a similar possibility of events
- Extreme BMI presents a risker sign







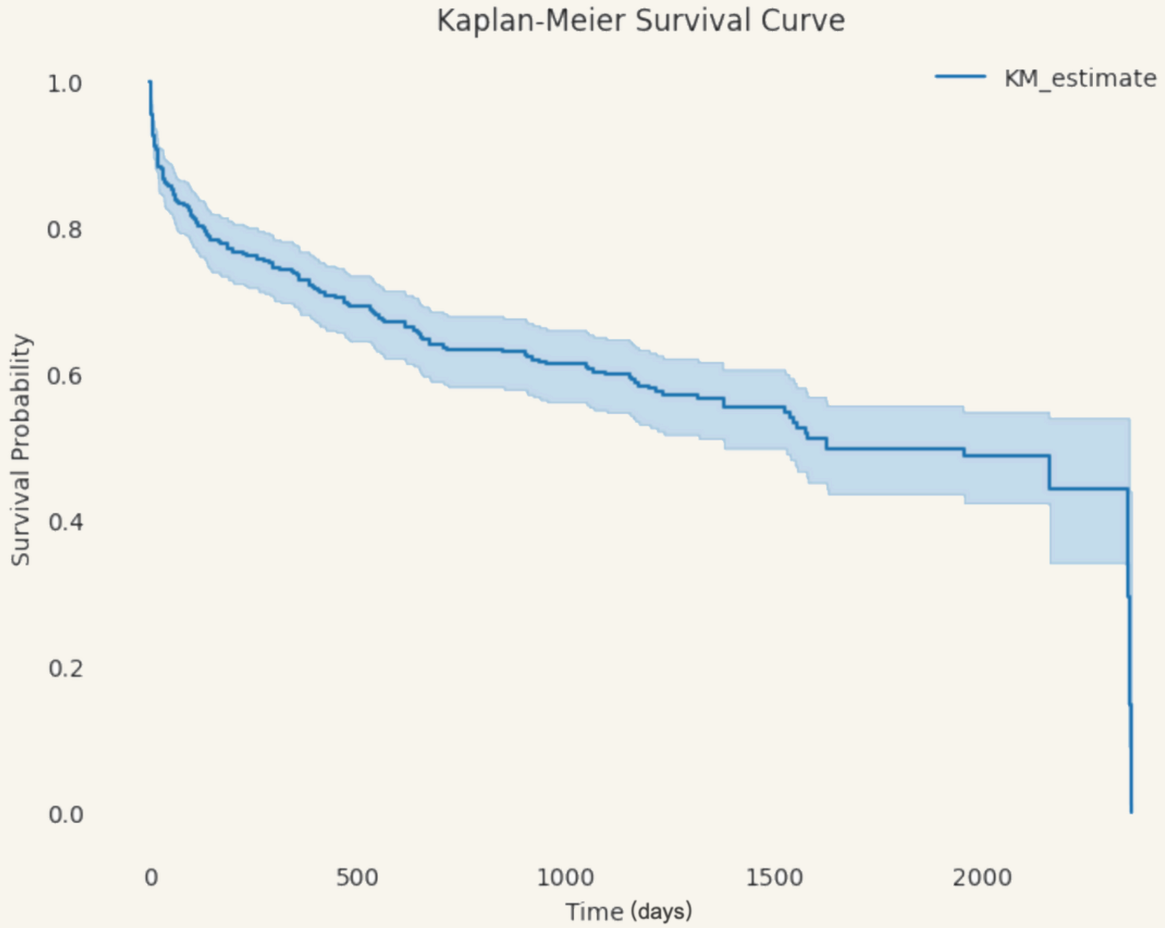
Ages:

- Samples are focused on the aged group

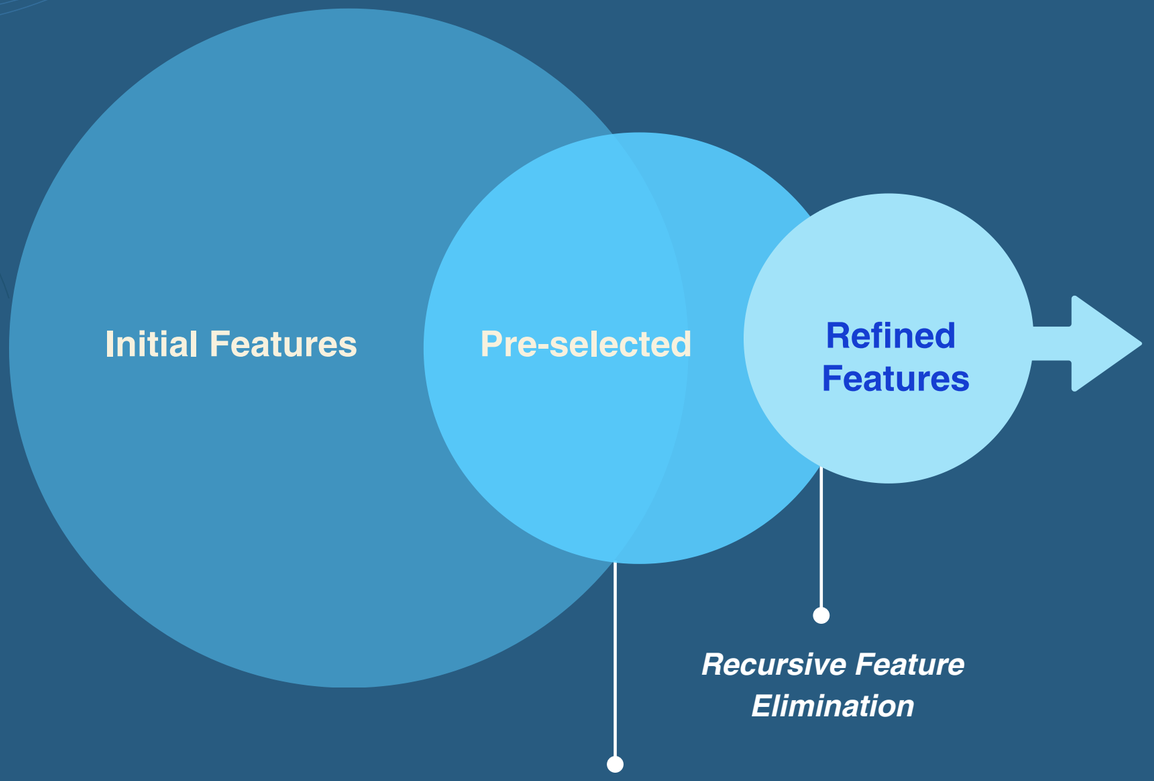
Survival Analysis

DATA EXPLORATION

A sharp decline in survival rates exhibits within three months. Then the drop alleviates until the end of the study well below 50%.



DATA PRE-PROCESSING



- *Create Mean Arterial Pressure(MAP)*
- *Convert 2 value categories into binary variables*
- *Convert race into dummy variables*

- age
- av3
- bmi
- chf
- gender
- hr
- mitype
- sho

COX-PH MODEL

Model Training

Concordance Index (c-index) on training dataset: ***0.8251***

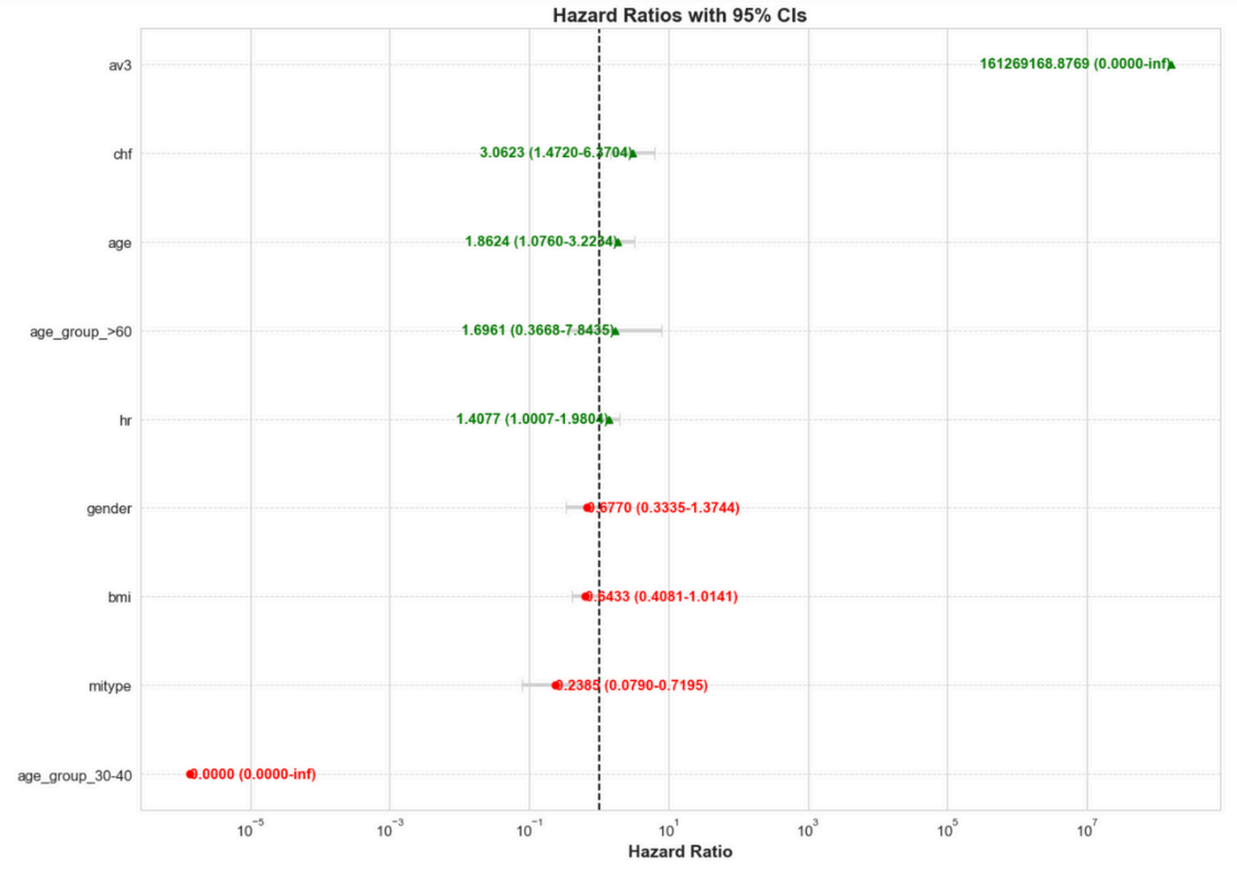
Model Validation

- **K-folds Cross Validation:** Average C-index across 5 folds: ***0.8200***
- **Bootstrapping with Cox-PH:**95.0 confidence interval of C-index: ***0.81 - 0.82***

Model Evaluation

Concordance Index (c-index) on test dataset: ***0.8251***

COX-PH MODEL INTERPRETATION



Key Takeaways from the Graph:

- **Age**: 86.24% ↑ risk/year (CI: [0.0733, 1.1704])
- **AV3**: Extremely high risk (CI: [-12,122.86, 12,160.66])
- **BMI**: 35.67% ↓ risk with ↑ BMI (CI: [-0.8962, 0.0140])
- **CHF**: 3x risk (CI: [0.3866, 1.8517])
- **Gender**: Females: 32.3% ↓ risk (CI: [-1.0983, 0.3181])
- **Heart rate**: 40.78% ↑ risk with ↑ rate (CI: [0.0007, 0.6833])
- **MI Type**: 76.15% ↓ risk for specific MI type (CI: [-2.5378, -0.3292])
- **Age 30-40**: Reduced hazard (CI: problematic)
- **Age >60**: 69.61% ↑ risk (CI: [-1.0031, 2.0597])

Business Implications:

- **Risk Management**: Prioritize older patients and CHF patients.
- **Resource Allocation**: Direct resources to high-risk groups.
- **Research**: Investigate problematic variables (e.g., AV3).

PYCOX MODEL
Implementation and Results

Configuration:
- Input Features: Based on dataset dimensions.
- Network Architecture:
- Layers: Input -> 128 -> 128 -> Output.
- Activation: ReLU.
- Regularization: Dropout (30%).
- Purpose: To model non-linearity and learn complex patterns.

Details:
- Optimizer: Adam.
- Batch Size: 256.
- Epochs: 20.

Steps:
- Convert tensors back to numpy for evaluation.
- Compute baseline hazards using a trained model.
- Predict survival curves for test data.
- Evaluate the model using C-index.

Key Result:
- C-index: 0.8183.
- Interpretation: Ability of the model to rank survival times.

Conclusion:
- Successfully implemented Cox's model with neural enhancements.
- The model's performance was evaluated using C-index.

04.

BUSINESS IMPLICATIONS & CONCLUSIONS

BUSINESS IMPLICATIONS

- By identifying key risk factors and providing actionable insights, our research contributes to enhancing patient care, optimizing resources, improving financial forecasting, and increasing federal reimbursements.
- The findings pave the way for a more personalized, efficient, and financially sustainable healthcare system, aligning with the dual purpose of health insurance organizations.
- The insights gained from this project can be instrumental in shaping future healthcare policies, strategies, and interventions, ultimately leading to improved quality of care and well-being for health plan members.

CONCLUSIONS

Enhanced Preventive Care and Improved Quality of Care

- By identifying high-risk members, health plans can manage conditions more effectively, potentially slowing disease progression and improving quality of life.
- Findings from the Cox-PH model, such as the significant risk associated with depression or gender, can guide these personalized interventions.

Profitability and Resource Optimization

- Implementing preventive measures and disease management programs can reduce high-cost care like hospitalizations and emergency room visits.
- Insights from the model, such as the importance of regular outpatient visits and medication compliance, can guide resource allocation, improving efficiency and reducing costs.

Financial Forecasting and Improved Risk Stratification

- By predicting adverse events, insurers can estimate future healthcare costs and inform pricing strategies.
- The insights into factors like race and inpatient admissions can help in proactive risk identification, potentially reducing the Medical Loss Ratio (MLR), a critical metric for health insurers.

THANK YOU

August 2023