Predicting Risk of Heart Disease Using Non-Typical Risk Factors

MGT 6203

Michael McCarthy, Member 2, Member 3

Georgia Tech

Introduction & Problem Statement

- Background
  - According to the American Heart Association, since 1921 heart disease remains the leading cause of death in individuals older than 20 years of age. For more than 100 years, the research and studies continue in attempt to decline rates. [1]
  - American's Heart Association Life's Essential Eight an outline to aid an individual to decrease their chances of getting heart disease.
- Problem Statement
  - We are trying to identify any non-common risk factors that hold similar predictive degree as commonly known risk factors of heart disease. Identifying these uncommon factors can provide a better insight of the individual's health and their chance of being diagnosed with heart disease.
  - Heart disease advances to its later stages it becomes more costly and harder to treat.[2] Using a well-rounded predictive model can lead to early invention helping the individual make a change before it is too late.
- Objectives
  - Main objective develop a predictive model to provide efficiency and reliable in heart disease factors
  - Subsidiary objectives include key risk factor identification, visualizations and reports, and actionable recommendations.

1. American Heart Association. (2024, January 24. *2024 Heart Disease and Stroke Statistics: A Report of US and Global Data from the American Heart Association.* Ahajournals.org. https://www.ahajournals.org/doi/full/10.1161/CIR.0000000000001209
2. Penn Medicine. (2022, July 18). *Heart Failure Classification - Stages of Heart Failure and Their Treatments*. Pennmedicine.org. https://www.pennmedicine.org/updates/blogs/heart-and-vascular-blog/2022/july/heart-failure-classification--stages-of-heart-failure-and-their-treatments

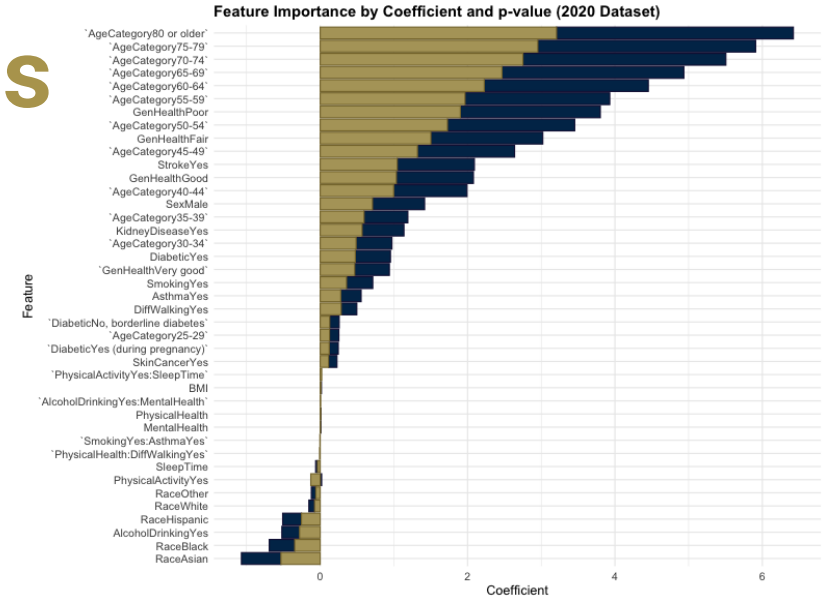
Georgia Tech

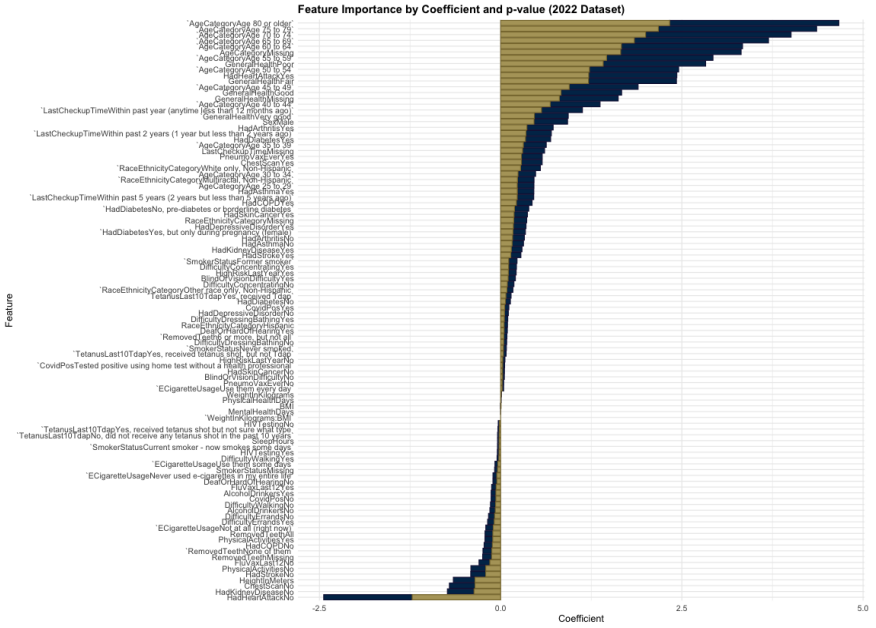
Key Insights & Findings

- **Best Model** - The "Full" Logistic Regression Model, using all variables, performs best with 5-fold Cross-Validation.

	2020		2022	
Measure	**Full Model**	**Model_C**	**Full Model**	**Model_C**
AIC	145,226	145,189	134,967	134,959
AUC	0.8407	0.8409	0.8952	0.8952
Accuracy	0.9159	0.9159	0.9453	0.9454
Sensitivity	0.9916	0.9917	0.9879	0.9879
Specificity	0.1068	0.1056	0.2807	0.2810

- **Minimal Gains:** No significant performance improvement from splitting variables or using "Non-Standard" predictors.
  - See visual from above
- **Feature Selection Impact** - Interaction variables and feature exclusions did not improve model performance.
  - There was similar feature importance between models and datasets.
- **Comprehensive Prediction** - Using all variables ensures the most comprehensive and simple risk prediction.





Georgia Tech

Business Impact & Implications

- Business Impact
  - With an efficient predictive model, it can lower the cost for the business and patients as well. The model can also provide a higher quality of life and care for the individuals.
  - Maintain a healthy and happy network of individuals thus retaining members and increasing the size of the network
- Implications
  - Collect the most data from individuals by obtaining a very detailed medical history
  - Store the data for easy access for providers and departments that use the data frequently
  - Develop a system to enforce and ensure individuals are remaining compliant with medications, diets, or wellness visits
  - Provide and offer educational health classes

Recommendations

- Develop Robust Data Infrastructure
  - **Support Growth**: Ensure the system can handle expanding data and new variables.
- Enhance Data Collection Methods
  - **Standardize Processes**: Consistent data collection reduces errors and simplifies analysis.
- Invest in Predictive Modeling
  - **Enable Early Intervention**: Use models to identify high-risk patients early and prevent issues.
- Implement Notification System for Physicians
  - **Quick Alerts**: Automate notifications to help physicians act on high-risk patients promptly.
- Ensure Data Governance & Compliance
  - **Ensure Security**: Follow regulations like HIPAA to protect patient data and maintain quality.

Implementation Plan

High Level Plan

- **Develop Data Infrastructure**
  - Build or upgrade a scalable database for storing patient data.
- **Streamline Data Collection**
  - Standardize and digitize patient surveys and health marker collection.
- **Create Predictive Models**
  - Develop models for assessing heart disease risk using current data.
- **Launch Notification System**
  - Integrate model outputs into physician workflows with alerts for high-risk patients.
- **Ensure Compliance**
  - Establish data governance protocols and ensure HIPAA compliance.

Resources and Support

- **Technical Support**
  - Data engineers for infrastructure upgrades and model implementation.
- **Healthcare Staff**
  - Training for physicians on using risk alerts.
- **Financial Resources**
  - Budget for database upgrades, predictive model development, and compliance tools.
- **Legal Expertise**
  - Ensure all data handling aligns with HIPAA and other regulations.

Timeline

- **0-3 Months**
  - Audit current infrastructure and data collection processes.
  - Begin database design and gather resources for upgrades.
- **4-6 Months**
  - Implement database upgrades and finalize data collection methods.
  - Begin feasibility analysis for predictive modeling (in-house vs. contracting).
- **7-18 Months**
  - Develop predictive models with ongoing evaluations for accuracy and reliability.
  - If contracting, identify and onboard external partners within the first 6 months of this phase.
- **19-24 Months**:
  - Test and refine the notification system to integrate risk alerts.
  - Final rollout of the complete system, including training for healthcare staff and ongoing compliance monitoring.

Georgia Tech

Challenges & Considerations

Potential Challenges

- **Technical Debt:** Data pipeline often require specialized tools and skillsets for data transformation.
  - Design & implementation of data workflow
  - Updates & maintenance
  - Cloud storage fees and compute power

- **Data Quality:** Real world data is often messy and with bias and ambiguity
  - Bias due to poorly worded survey questions or non-random sampling
  - Missing or incomplete information
  - Duplicate information
  - Different nomenclatures e.g., date format

Mitigation Strategies

- Roche's Maximum which state, "Data should be transformed as far upstream as possible, and as far downstream as necessary" (Roche, 2023).
- Careful selection of tool for data transformation and data storage.
- Data governance to address business rules and logic, with special attention to data privacy rules and regulations. I.e., HIPPA

Roche, M. (2023, April 17). *Roche's Maxim of Data Transformation*. BI Polar. https://ssbipolar.com/2021/05/31/roches-maxim/

Georgia Tech

Future Directions

Further Analysis

- Missing observations from crucial variables e.g., HadAngina, HadHeartAttack, SmokerStatus, and BMI
- Bias towards mid-west and eastern region of the United States
- Bias towards White only, Non-Hispanic

Long-term Improvements

- Data Accessibility, Quality, and Governance
  - Invest in developing or improving robust data pipelines with clear and concise business logic that facilitates data quality, efficiency, and flexibility.
  - Employ simple random sample techniques such as lottery or random number to mitigate bias.
  - Review and ensure questions are clear and concise to mitigate missing observations.
  - Offer incentives to fully complete the survey.
  - Implement data quality control measure for early detection of inconsistencies and completeness.

Conclusion & Summary

Summary

- With our research we aimed to provide an optimal model that can predict unspecified risk factors of heart disease.
- Preparing and cleaning the data: VIF, Cook's Distance, and Imputation.
- Optimal model: Logistic Regression with 5-Fold Cross-Validation.
- Key metrics for optimal model: Confusion Matrix, ROC & AUC, and AIC.
- Common high-risk factors included Had Heart Attack, General Health, Age, and BMI.
- Uncommon high-risk factors included Sex, Sleep Hours, and Removed Teeth.

Closing

- We believe that these insights were only possible because of the extensive collection of information and it is for that reason that we recommend developing a more robust data infrastructure that entails enhancing data collection methods, invest in predictive modeling, creating a notification system for physicians, and implement data governance specific to healthcare data.

Georgia Tech