Heart Attack Policyholder Data Analysis and Initiative Proposal

Rachel Goldsbury

SNHU, DAT 510

September 3, 2023

Professor Kuchibhotla

Heart Attack Patient Data Analysis and Initiative

Introduction

The health insurance program director, Sonia has the goal of figuring out which ones of her policyholders are at a high risk for suffering from a second heart attack. We do not have many details about the company or Sonia except that she has in the past used data mining for another initiative and now she is ready to tackle a new one with the right funding in place. We will focus on Heart attack risk management. There's a lot that can be done to lower the risk of heart attack victims suffering from a second heart attack and it's essential to minimize this risk to extend the lives of people, as well as save money for the company from paying out on claims and be able to lower the premiums for patients.

We want to provide the patients with what they need to improve their lifestyle choices, including; weight and stress management. The goal of this data analysis project is to pinpoint the correct training data that meets the criteria of accurately predicting high risk patients and implement a risk management strategy. The pros of this analysis is that patients will lower their risk of a second heart attack and extend the lives of real people so they can spend more time with the people they love in life. Additionally, we want to figure out which patients are a good fit for which programs. We will be using logistic regression as a predictor tool for determining the likelihood of this event.

We have access to the company's medical claims database. This comes with benefits and limitations but overall the current usage is appropriate for our goal. We can always add more data points to our analysis if needed to increase the accuracy. The information in our medical claims database includes age, marital status, gender, weight-category, cholesterol, stress management, trait anxiety.

Gender, age and weight, cholesterol and anxiety can all be factors in heart attacks, we already know this. This information is good to have but limitations exist where we don't have information on their lifestyles. What they eat, do they work out and family history of heart attacks. In terms of stress and anxiety measuring these can be difficult, maybe a better measure would be blood pressure. These are all also contributing factors and it is to our disadvantage that we don't have this data. On the other hand the data we do have may already be enough to predict the issue with enough accuracy so that we don't actually need any of the other data. We will need to see how accurate our prediction is with the information we have and if it is not accurate enough (a number which will need to be determined), then we need to gather more data of risk factors. It would most likely be beneficial to include this to benefit the organization.

I think bringing in additional data is going to be important, more on the lifestyle choices of people and understanding if they work out, do they eat processed foods, etc will be a data need to make this project as accurate as possible.

Proposal

Section A: Goals

The goal of this analytic initiative is to identify policyholders who are at a high risk for suffering from a second heart attack and provide them with community education resources to help lower their risk and extend their lives. This aligns with the organizational goal of improving health, lowering premiums for patients and reducing costs for the company. Success will be measured by the accuracy of this prediction model in identifying high-risk patients and the subsequent reduction in second heart attacks amongst these individuals, ultimately significantly decreasing the incidents of second heart attacks and overall patient outcome.

Section B: Data Analytic Life Cycle

Problem Definition: Clearly defining the problem of identifying high-risk patients for a second heart attack and the objectives of the analysis.

Data Collection: Gather relevant data from the company's medical claims database, including age, marital status, gender, weight-category, cholesterol, stress management, and trait anxiety.

Data Preparation: Clean and preprocess the data, addressing missing values and outliers, and possibly adding more data points if necessary.

Exploratory Data Analysis: Perform exploratory analysis to gain insights into the relationships between variables and potential patterns related to second heart attacks.

Model Development: Use logistic regression as a predictive tool to develop a model that can accurately predict high-risk patients based on the available data.

Model Evaluation: Assess the model's performance using appropriate evaluation metrics such as precision, recall, and F1-score. Fine-tune the model if needed.

Deployment and Action: Deploy the model to identify high-risk patients and provide them with targeted community resources to reduce their risk of a second heart attack.

Monitoring and Maintenance: Continuously monitor the model's performance and update it as necessary to ensure its accuracy and effectiveness over time.

Section C: Value of the Life Cycle

The data analytic life cycle plays an important role in this initiative by ensuring accurate predictability, performance and quality of the results. Through each stage of the life cycle we will be able to refine our predictive model to enhance its accuracy. By analyzing the data and improving the model, we can predict high-risk patients more effectively leading to better allocation of resources for interventions and improved patient

outcomes. The life cycle also ensures that data quality issues are addressed early on, contributing to the reliability of the model's predictive outcomes.

Section D: Data

The existing data from the company's medical claims database provides valuable information. The information includes data on age, marital status, gender, weight, cholesterol and stress management/ anxiety levels, cortisol levels. These factors are known to contribute to heart attacks, however limitations exist due to the absence of lifestyle data such as diet, exercise and family history. This could further enhance the accuracy of the model. While the available data can provide potentially accurate predictions, lifestyle data may significantly improve the models performance. Gathering this additional data should be considered while ensuring data security is maintained to protect the patients personal information.

Section E: Tool Applicability to Initiative

R Studio is a comprehensive environment for data analysis, visualization, and statistical modeling that aligns well with our goals. R Studio's integration with R packages and libraries allows us to perform data manipulation, exploratory data analysis, and model development seamlessly. Its ability to handle structured data and statistical functions has been instrumental in the development of our logistic regression model.

Section F: Tool Applicability

R Studio is well-suited to handle the structured data we currently possess, which includes variables such as age, marital status, gender, weight-category, cholesterol, stress management, and trait anxiety. Its strengths in statistical analysis and visualization make it suitable for uncovering insights and patterns within this data. It will also be suitable if we end up gathering additional lifestyle data.

Section G: Tool Recommendations

Our focus will be in R and RStudio, After some research I can recommend the following R packages to enhance our data analytics initiative. These recommendations align well with our current toolset and goals, allowing us to expand our capabilities within the R environment, improve prediction accuracy, and effectively communicate insights to stakeholders.

Caret Package: With Caret, we can easily experiment with various algorithms, perform feature selection, and optimize model parameters. This package will allow us to extend our predictive modeling capabilities and potentially handle more complex data sources, contributing to the accuracy of our predictions.

Shiny Package: The Shiny package enables us to create interactive web applications directly from R. This tool could enhance our ability to present our analysis and predictions to stakeholders in an engaging and user-friendly manner. By developing interactive dashboards using Shiny, we can provide a more intuitive way for decision-makers to explore insights and make informed choices based on our predictions.

Conclusion

Applying data analytics to our company, especially within the scope of our initiative to identify high-risk patients for a second heart attack, holds immense value. Our analysis and the proposed initiative are positioned to benefit the organization in several key ways: patient care, cost savings and Improved Business Efficiency. By accurately identifying high-risk patients, we enable the company to provide targeted interventions and resources to those who need it the most. This not only improves patient outcomes and their quality of life but also reduces the outcome of future costly medical procedures associated with second heart attacks. This will result in a direct cost savings benefit to the organization. The model we will develop will additionally assist in optimization of resources to those who need them the most and lowers the unnecessary spending on low-risk individuals.

Insights

Our analysis and exploration of data analytic tools has yielded valuable insights that have the potential to shape the company's future and my growth as an analytics professional. It allows us to explore a data-driven decision approach made by basing intervention on predictive models rather than generalized approaches. The company can strategically manage resources and deliver continuous improvement. Regular iteration of the data analytic life cycle will maintain accuracy and relevance over time.

Communication to stakeholder and Presentation Considerations: To effectively communicate these results to stakeholders copies of this as well as an executive summary of the key findings will be provided. Actionable recommendations will be

clearly outlined with predictive models, resource allocation and intervention strategies.

Visualization: Techniques we will use are line charts to keep track of our progress over time, bar charts to compare and analyze categorical data, heat maps to show correlation between variables.

Preliminary Training Set Data Analytics

training_data_analytics

variable	average	mode	median	st dev	0	1	2
Age	62.97826087	66.0	63.0	7.853092181			
marital status		2.0					
gender		1.0					
weight cat		1.0					
cholesterol	177.3913043		172.0	32.26270695			
Stress mgmt		0.0					
trait anxiety	55.47445255	60.0	55.0	12.40972122			
2nd heart attack (yes)	68.0						
2nd heart attack (no)	70.0						
marital status					33.0	50.0	56.0
gender					53.0	86.0	0.0
Stress mgmt					73.0	61.0	0.0

Preliminary Patient Data Analytics

patient_data_analytics

variable	average	mode	median	st dev	0	1	2
Age	62.93188406	66.0	63.0	7.898650337			
Marital Status		2.0					
gender		1.0					
weight cat		1.0					
cholesterol	178.2652174	173.0	173.0	32.28902641			
stress mgmt		0.0					
trait anxiety	55.43478261	60.0	55.0	12.33715789			
marital status					40.0	250.0	280.0
gender					260.0	430.0	0.0
stress mgmt					375.0	315.0	0.0

Citations

1. North, M. (2012). *Data mining for the masses*.