BA 723

Business Analytics Capstone

(A Data Governance Project)

“Building Trustworthy Heart Disease Prediction: A Community-Driven Approach”

Prepared

By:

Jason S. Yap

301293413

Introduction

This document outlines the importance of effective model validation and governance for ensuring reliable heart disease prediction models.

Building a Robust Model:

- **Rigorous Validation:** We employ comprehensive validation techniques to prevent flawed models, including testing, model construction examination, and data verification. (Inspired by Domino Data Lab)
- **Model Governance:** Established rules, protocols, and access controls safeguard the model's lifecycle and address potential issues like performance drift. (Inspired by Analytics Insight)
- **Transparency and Fairness:** Mitigating biases and ensuring transparency are key aspects of our approach. (Inspired by Yields.io)

Inspired by BRFSS, Localized for Action:

- **Project Goal:** Develop a localized, digital survey system similar to the US-based Behavioral Risk Factor Surveillance System (BRFSS) to identify individuals at risk of heart disease.
- **Data Foundation:** The initial model utilizes a heart disease indicator dataset from BRFSS.
- **Community Focus:** This system is designed for use by communities, non-profit organizations, and healthcare institutions.

Transitioning to User-Owned Data:

This document details the steps involved when communities and organizations upload their own survey datasets for prediction.

Methodology

To replicate the predictive model, the following steps must be completed:

1. Data Cleaning and Preparation

- Handle missing values through imputation as needed.

- Address inconsistencies or outliers using capping and flooring if necessary.
- Convert categorical variables into a numerical format using one-hot or label encoding.
- Normalize or standardize numerical features to ensure consistent scale.

2. Exploratory Data Analysis

- Calculate descriptive statistics (mean, median, standard deviation) for each lifestyle factor.
- Visualize feature distributions and relationships with heart disease risk using skewness, histograms, and box plots.

3. Correlation and Feature Selection

- Conduct correlation analysis to identify potential relationships between features.
- Employ a Random Forest classifier to identify significant features for modeling.

4. Predictive Modeling

- Build multiple models (Decision Trees, Logistic Regression, Neural Networks) using selected features.
- Evaluate model performance using ROC AUC, accuracy, recall, precision, and F1 score.
- Apply cross-validation to assess model robustness and generalizability.

5. Model Selection and Interpretation

- Select the best-performing model based on evaluation metrics.
- Interpret model results to identify key indicators of heart disease risk.

6. Recommendations and Insights

- Utilize insights from the best indicators to inform heart disease risk identification and prevention.
- Create visualizations to effectively communicate findings.
- Prepare a comprehensive report summarizing the analysis, results, and recommendations.

1.0. Variable Level Monitoring

Variable-level monitoring is crucial for maintaining model health and data integrity over time. By tracking changes in input variable characteristics, it enables early detection of

issues like data drift, quality problems, and model performance degradation. This proactive approach helps prevent biases, compliance violations, and data inconsistencies. Furthermore, by identifying the root causes of model behavior changes, it supports efficient troubleshooting and ensures the model remains aligned with its intended purpose.

1.1. Model Build Variable Level Statistics Variable Distributions

Table 1: Descriptive Statistics

Descriptive Statistics	Type	count	mean	std	min	25%	50%	75%	max
HeartDiseaseorAttack	Binary	47786	0.5	0.500005	0	0	0.5	1	1
HighBP	Binary	47786	0.57002	0.495078	0	0	1	1	1
HighChol	Binary	47786	0.54769	0.497726	0	0	1	1	1
CholCheck	Binary	47786	0.97393	0.159359	0	1	1	1	1
Smoker	Binary	47786	0.52042	0.499588	0	0	1	1	1
Stroke	Binary	47786	0.09612	0.294753	0	0	0	0	1
Diabetes	Binary	47786	0.47367	0.836709	0	0	0	0	2
PhysActivity	Binary	47786	0.70713	0.455083	0	0	1	1	1
Fruits	Binary	47786	0.61876	0.485697	0	0	1	1	1
Veggies	Binary	47786	0.79126	0.406415	0	1	1	1	1
HvyAlcoholConsump	Binary	47786	0.04784	0.213426	0	0	0	0	1
AnyHealthcare	Binary	47786	0.95605	0.204977	0	1	1	1	1
NoDocbcCost	Binary	47786	0.09574	0.294237	0	0	0	0	1
GenHlth	Binary	47786	2.89014	1.15812	1	2	3	4	5
MentHlth	Open Cont.	47786	3.85965	8.298438	0	0	0	2	30
PhysHlth	Open Cont.	47786	6.42148	10.5365	0	0	0	7	30
DiffWalk	Binary	47786	0.27797	0.448002	0	0	0	1	1
Sex	Binary	47786	0.50128	0.500004	0	0	1	1	1
Age	Category	47786	8.97583	2.907543	1	7	9	11	13
Education	Category	47786	4.91791	1.030567	1	4	5	6	6

Income	Category	47786	5.65364	2.172331	1	4	6	8	8
BMI_Numeric	Categorized	47786	2.32539	1.150381	0	1	2	3	5

The descriptive statistic table above shows that many variables are binary (0 or 1), indicating the presence or absence of a condition (e.g., HeartDiseaseorAttack, HighBP, Smoker). Others, like GenHlth, MentHith, and PhysHith, are ordinal with limited categories. Lastly, numerical variables such as Age, Education, Income, and BMI Numeric are categorized variables with varying ranges and distributions.

Potential Issues

- **Outliers:** Variables like MentHith and PhysHith have maximum values significantly higher than the 75th percentile, suggesting potential outliers that could impact analysis. However, in the case of the BRFSS heartrisk disease dataset, the variables were retained as it is, due to being continuous variables with varying ranges and distributions, which makes it difficult to categorize. When outliers are present, one can perform the following strategies:
    - **Removing Outliers**:
        - **Simple Removal**: If outliers are due to data errors, they can be removed.
        - **Domain Knowledge**: Use domain knowledge to decide if extreme values should be excluded.
    - **Transforming Data**:
        - **Log Transformation**: Apply (to open continuous variables only) a log transformation to reduce the impact of high values.
    - **Imputation**:
        - **Replace with Mean/Median**: Replace outliers with the mean or median of the data (Applicable to continuous variables only).

- Data Imbalance: Some binary variables (e.g., Stroke, HvyAlcoholConsump, NoDocbcCost) have a low proportion of positive cases, which might affect model performance if not addressed appropriately. If the binary variables is too low, one

can perform upscaling or the SMOT technique. However, for the current dataset its unnecessary since the general population is healthy.

1.2. Acceptable Ranges

Establishing clear and realistic value ranges for each input variable is crucial for data quality and model accuracy. Defining acceptable boundaries prevents anomalous or outlier data from corrupting the model's training process. This ensures the model learns from relevant and meaningful data, ultimately improving its predictive capabilities.
Currently, most variables in the dataset are binary. Others, like MentHlth and PhysHlth, are open-ended and continuous. Lastly, numerical variables such as GenHlth, Age, Education, Income, and BMI Numeric are categorized as ordinal variables with defined ranges and distributions according to the BRFSS-CDC's standards.

Categorized Ordinal Variables

Once the predictive model is used by different institutions, they may approach their survey questionnaires differently to gain new insights. They will likely prefer some, if not most, of the categorized ordinal variables as open continuous. In such cases, the institutions should:

1. Explore the distribution by calculating skewness and visualizing the data with boxplots and histograms. This analysis helps determine appropriate ranges or categories for the target market or community.
2. Feature engineer the determined categories and convert them into numerical ordinal variables for easier and more concise modeling.

In summary, appropriate value ranges for input variables should align with the specific characteristics of the target market or community. On the other hand, to minimize bias and subjectivity, these ranges should adhere to established government classifications

whenever possible. For instance, BMI categories can follow the guidelines set forth by the Government of Canada.

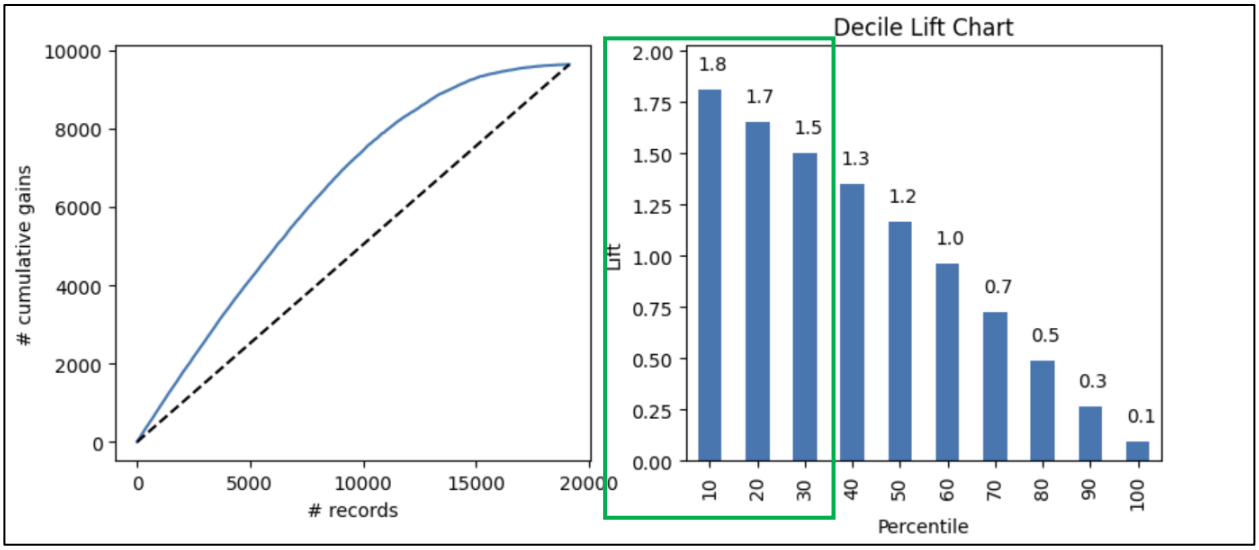
Health Risk Classification According to Body Mass Index (BMI)

Classification	BMI Category (kg/m^2)	Risk of developing health problems
Underweight	< 18.5	Increased
Normal Weight	18.5 - 24.9	Least
Overweight	25.0 - 29.9	Increased
Obese class I	30.0 - 34.9	High
Obese class II	35.0 - 39.9	Very high
Obese class III	>= 40.0	Extremely high

The image illustrates the BMI categories defined by the Government of Canada. These categories were applied to the heart disease dataset, encompassing individuals with BMIs ranging from 18.5 to 40 or higher.

1.5. Variable Drift Monitoring Tolerance

To maintain model accuracy and relevance, it's essential to set clear thresholds for detecting variable drift. When the statistical properties of input data change significantly, it can compromise model performance. By closely monitoring these shifts and implementing timely adjustments like retraining or updating, we can safeguard the model's predictive capabilities.



The proponent employs a tiered approach to managing feature drift. Critical features have a stricter drift tolerance of 10% deviation from their training data mean, while less critical features allow for a 40% deviation. Before automatically rebuilding the model when drift thresholds are exceeded, the proponent rigorously assesses model health using AUC-ROC and F1-score metrics. Subsequent actions are determined based on the identified risk level.

Model Monitoring, Health & Stability

2.0. Initial Model Fit Statistics

To assess model performance and ensure its ongoing effectiveness in driving business decisions, the model risk management framework mandates regular qualitative and quantitative evaluations. To monitor model health and stability, the framework utilizes AUC-ROC, Accuracy, Recall, Precision, F1 Score metrics as key performance indicators and benchmarks.

- The ROC curve visually depicts a binary classifier's performance by charting its true positive rate against its false positive rate across various classification thresholds. The AUC metric quantifies this overall performance, indicating the model's ability to distinguish between classes.

- Accuracy is a straightforward metric measuring the proportion of correct predictions among all instances. While effective for balanced datasets, it can be misleading when dealing with imbalanced classes.
- Recall is crucial since it prioritizes identifying all positive instances. This metric is highlights domains like medical diagnosis where missed positive cases can have severe consequences.
- Precision is essential in determining a good balance between accuracy and recall score.
- The F1 score balances precision and recall, providing a comprehensive evaluation metric. It is particularly valuable when dealing with imbalanced datasets where both false positives and false negatives are significant.

In summary, these metrics complement each other to ensure accurate predictive modeling.

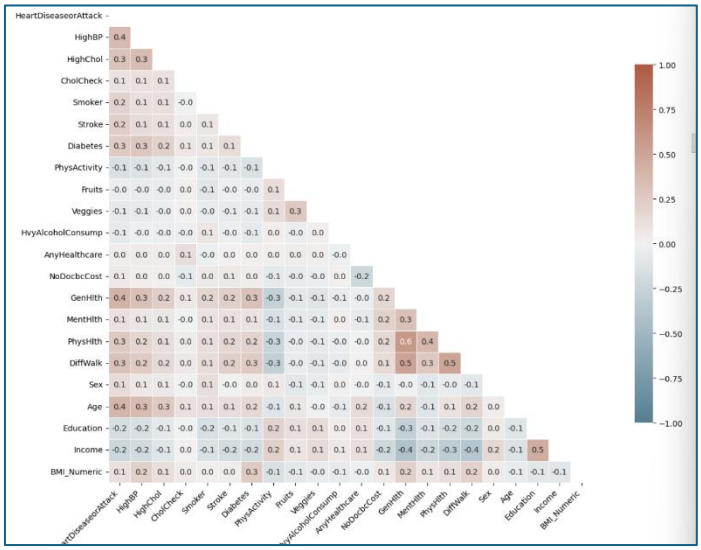
2.1. Model Comparison and Acceptable Threshold

Model Name	ROC AUC	Accuracy	Recall	Precision	F1 Score
Random Forest	0.7794	0.7180	0.7478	0.7085	0.7276
Full ClassTree	0.6672	0.6595	0.6293	0.6734	0.6506
Decision Tree	0.6671	0.6595	0.6284	0.6736	0.6502
Small Class Tree	0.7479	0.6805	0.7496	0.6613	0.7027
Small ClassTree3	0.7876	0.7294	0.7653	0.7167	0.7402
SmallClassTree5	0.8134	0.7445	0.7988	0.7230	0.7590
Grid Class Tree	0.8074	0.7425	0.8177	0.7131	0.7619
Logistic Regression	0.8249	0.7531	0.7874	0.7394	0.7626
Backward Elimination	0.8249	0.7531	0.7874	0.7394	0.7626
Forward Selection	0.8249	0.7531	0.7874	0.7394	0.7626
Stepwise Selection:	0.8249	0.7531	0.7874	0.7394	0.7626
Neural Network	**0.8255**	**0.7520**	**0.8045**	**0.7304**	**0.7657**
Acceptable Threshold	**0.7829**	**0.7249**	**0.7576**	**0.7131**	**0.7340**

The model with the highest score for the key metric indicator is highlighted in green. The average score across all models serves as the acceptable threshold and is indicated in yellow.

2.2. Heat Map Correlation

To optimize key indicator scores, conducting a correlation heatmap analysis is essential after exploration or before modeling. It visually depicts relationships between variables. They are essential for identifying strong correlations, detecting redundant information (multicollinearity), and selecting the most informative features for model building. Detecting redundant information (multicollinearity) is crucial as it will be difficult to interpret the results later.



The image displays a correlation heatmap where a threshold of 0.7 is considered the maximum acceptable correlation. Since the highest correlation in this example is 0.6, it indicates a moderately correlated relationship which is generally acceptable and can be used for modeling. However, if a correlation exceeds 0.7, further investigation and potential actions are necessary.

Dealing with Correlated Variables

When variables are too closely related (highly correlated), it can cause problems in building accurate models. Here's what data analysts can do:

- **Remove one variable:** If two variables provide similar information, one can be removed without losing much data.
- **Combine variables:** Sometimes, related variables can be merged into a single variable to simplify the data.

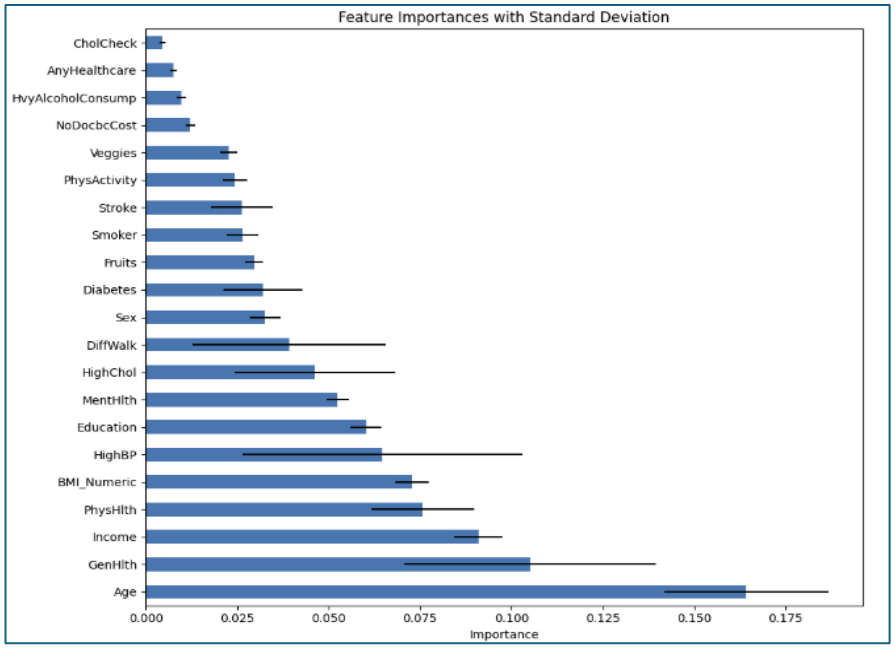
- **Dimensionality reduction:** Techniques like PCA can transform many variables into fewer, less correlated ones.
- **Penalize strong relationships:** Methods like Ridge and Lasso regression can handle correlated variables by downplaying their importance.
- **Measure correlation strength:** The Variance Inflation Factor (VIF) helps identify how much one variable is influenced by others. High VIF indicates a problem.
- **Adjust data:** Centering data (subtracting the average) can sometimes help, especially in complex models.
- **Alternative modeling:** Partial Least Squares Regression is a method designed for situations with many correlated variables.

2.3. Random Forest

After addressing multicollinearity by removing highly correlated variables or using techniques like PCA, the remaining features can be used for model building. This helps identify the most important features before training models like Random Forests, Decision Trees, Logistic Regressions, and Neural Networks.

Random Forest Is a powerful and versatile ensemble learning method used for both classification and regression tasks. It builds multiple decision trees during training and merges them to improve accuracy and control overfitting. Random Forest is used for random feature selection in this case.

	feature	importance	std
17	Age	0.164292	0.022443
12	GenHlth	0.105131	0.034418
19	Income	0.090959	0.006581
14	PhysHlth	0.075790	0.014060
20	BMI_Numeric	0.072842	0.004563
0	HighBP	0.064715	0.038310
18	Education	0.060235	0.004254
13	MentHlth	0.052450	0.003021
1	HighChol	0.046289	0.021898
15	DiffWalk	0.039326	0.026379
16	Sex	0.032603	0.004288



The feature importance plot, generated from Random Forest, identifies nine key features out of 22 for subsequent modeling. These significant features will be used to build Decision Trees, Logistic Regression, and Neural Network models.

Note: While percentage-based cutoffs (e.g., top 20%) or statistical methods like chi-square or ANOVA are generally recommended for determining optimal feature selection thresholds, this analysis employs a fixed value of 5% of the highest importance to explore potential insights within the decision tree model and odds ratio results.

2.4. Performance Over Time

Given the critical nature of heart disease prediction, rigorous and frequent monitoring of model performance is imperative. Track model performance evolution by conducting regular assessments and documenting significant changes in key metrics such as accuracy, sensitivity, specificity, and AUC. This proactive approach ensures the model's effectiveness in identifying heart disease over time.

3.0 Risk Tiering

Risk Tiering is a crucial aspect of model governance, especially for models with significant decision-making implications, such as the heartdisease_indicator_balance model. This framework categorizes models based on risk factors like performance, data quality, complexity, and impact. Its primary goal is to proactively manage model-related risks and ensure optimal performance.

3.1 Risk Assessment Criteria

- **Performance Metrics:** Evaluate key performance indicators (accuracy, precision, recall, F1-score, AUC) to identify performance degradation.
- **Data Integrity:** Assess data quality and consistency to detect potential issues impacting model reliability.
- **Model Complexity:** Consider model intricacy (variables, algorithm) as it influences interpretability and maintenance.
- **Model Impact:** Evaluate the model's influence on decision-making processes. High-impact models necessitate closer monitoring.

3.2 Risk Tier Categories

- **Low Risk:** Consistent performance, minimal data issues, low impact (Drift is less than 40%). Routine monitoring is sufficient.
- **Medium Risk:** Performance or data quality concerns, moderate impact (Drift is between 50% to 80%). Increased monitoring and potential interventions.
- **High Risk:** Significant performance degradation, severe data issues, or high impact. Immediate investigation and potential model remediation (Drift is over 90%).

3.3 Action Plan per Risk Tier

- **Low Risk:** Quarterly monitoring, annual reviews, minimal documentation.
- **Medium Risk:** Monthly monitoring, quarterly reviews, detailed documentation, and stakeholder reports.

- **High Risk:** Weekly monitoring, immediate action, comprehensive documentation, frequent stakeholder communication.

3.4 Mitigation Strategies

- **Threshold Adjustment:** Optimize decision thresholds to balance false positives and negatives.
- **Model Refit:** Retrain the model with new data to adapt to changing conditions.
- **Model Rebuild:** Consider a complete model rebuild in case of severe performance issues or data changes.

3.5 Review and Approval Process

- **Review Frequency:** Align review frequency with risk tier (more frequent for high-risk models).
- **Approval Hierarchy:** Define approval paths based on risk level, involving relevant stakeholders.
- **Documentation:** Maintain comprehensive records of reviews, approvals, and actions taken.

References:

Behavioral Risk Factor Surveillance System. (2024). Annual Survey Data: https://www.cdc.gov/brfss/annual_data/annual_data.html

Government of Canada. (2024). Body Mass Index (BMI) Nomogram: https://www.canada.ca/en/health-canada/services/food-nutrition/healthy-eating/healthy-weights/canadian-guidelines-body-weight-classification-adults/body-mass-index-nomogram.html

Domino. (2021). The Importance of Machine Learning Model Validation and How It Works: https://domino.ai/blog/what-is-model-validation

MLOps and Model Governance. (2024). https://ml-ops.org/content/model-governance

Yields.IO. (2024). Model Risk Management: https://www.yields.io/

World Health Organization. (2024). Cardiovascular diseases: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1Kaggle. (2022). Heart Disease Health Indicators Dataset: https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset American Heart Association. (2024). Understand Your Risks to Prevent a Heart Attack: https://www.heart.org/en/health-topics/heart-attack/understand-your-risks-to-prevent-a-heart-attack#:~:text=An%20inactive%20lifestyle%20is%20a,blood%20pressure%20in%20some%20people. National Library of Medicine. (Circa). Risk Factors for Coronary Artery Disease: Historical Perspectives: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5686931/

Parent, D. (2023). Business Analytics and Insights notes. Centennial College.