Module 6 Final Project — Milestone 2: Final Report

ALY6040: Data Mining, Northeastern University

By: Group 2

Yash Gokhale
Yulin Liang
Aadira Anil Ramakrishnan
Mohit Kamble
Monika Gundecha

Professor Justin Grosz

6/24/24

Introduction

As data analysts for a health insurance company, we're leveraging the National Health and Nutrition Examination Survey (NHANES) dataset to address a critical challenge: predicting kidney failure risk among our clients. Kidney failure is a severe health condition with significant medical and financial implications. By developing accurate predictive models, we aim to enhance our risk assessment capabilities and tailor our insurance offerings more effectively. Our project focuses on two primary business questions:

1. How can we accurately identify individuals at high risk of kidney failure using available health data?
2. How can we optimize our insurance packages and pricing strategies based on these risk assessments?

By analyzing relevant variables from the NHANES dataset, we've developed predictive models to assess kidney failure risk. This data-driven approach will enable our company to make informed decisions about insurance coverage, implement targeted preventive measures, and potentially improve health outcomes for our clients. Our findings will not only guide the creation of specialized insurance packages but also contribute to more accurate premium calculations.

Data Cleaning

Our data mining project began with the acquisition of 6 datasets from the National Health and Nutrition Examination Survey (NHANES). These datasets covered a wide range of health-related information, including demographics, diet, medical examinations, laboratory results, medication use, and health history. Each dataset contained multiple variables, presenting us with a rich but complex data landscape. We then joined these datasets using the primary key of user ID, creating a comprehensive dataset for our analysis. The data cleaning process was approached with careful consideration, tailoring our strategies to the unique characteristics of each variable. Here's a detailed breakdown of our approach:

1. **Removing Columns with High Percentage of Null Values:** We identified columns with an excessive amount of missing data. Variables such as BPXDI4 (Systolic blood pressure reading 4), BPXCHR (60-second blood pressure), and DIQ280 (Last HbA1c level) contained more than 60% null values. Given the high proportion of missing data, these columns were removed from the dataframe to ensure the overall quality and reliability of our analysis.
2. **Treating Dietary Variables:** Variables related to dietary intake, including protein (DR1TPROT), saturated fat (DR1TSFAT), monounsaturated fat (DR1TMFAT), and polyunsaturated fat (DR1TPFAT), had approximately 9-10% missing values. For these variables, we employed an age group-specific median imputation strategy. This approach was chosen because dietary patterns can vary significantly across different age groups due to factors like lifestyle, cultural influences, and physiological changes associated with aging. By using age-group-specific medians, we ensured that the imputed values were more representative and accurate, accounting for these potential variations. We avoided mean

imputation for these variables because dietary data often exhibits skewed distributions, and means can be heavily influenced by outliers, potentially leading to less reliable imputations.

3. **Handling Income and Pulse Rate**: Variables such as total income (INDFMIN2) and pulse rate (BPXPLS) had relatively low percentages of missing values at 1.2% and 3.2%, respectively. These variables often follow normal distributions with some variation. For these, we replaced missing values with the median value of the respective column. The median was chosen as it provides a stable central value that is less affected by outliers or extreme values, making it a suitable choice for imputation. We opted against mean imputation because it can be heavily influenced by extreme values, especially in the presence of outliers, which could lead to inaccurate imputations.
4. **Addressing Biomarker Variables:** Biomarkers like serum creatinine (LBDSCRSI), serum albumin (LBDSALSI), blood urea nitrogen (LBXSBU, LBDSBUSI), and creatinine (LBXSCR) had 4-5% null values and exhibited skewed distributions. For these variables, we replaced missing values with the median value of the respective column. The median value is less affected by extreme values, which are common in biomarker data, providing a more representative measure for imputation compared to the mean. We avoided mean imputation for these variables because it can be heavily influenced by outliers, potentially leading to inaccuracies in the imputed values.
5. **Blood Pressure Variables:** Blood pressure variables BPXDI1, BPXDI2, and BPXDI3 (diastolic) and BPXSY1, BPXSY2, and BPXSY3 (systolic) had around 5-8% missing values. These missing values were imputed using the median value of the respective column, as they exhibited skewed distributions, and the median provides a robust central measure less affected by outliers. After cleaning, we created new columns named average diastolic blood pressure (BPXDI_avg) and average systolic blood pressure (BPXSY_avg) using the average of the three respective readings for each measure.
6. **Medication Usage (OSQ130): The** variable OSQ130, indicating whether an individual has taken prednisone or cortisone for an extended period, posed a significant challenge with over 30% of the values being missing. Simple substitution methods like mode imputation or machine learning techniques such as K-Nearest Neighbors (KNN) and Random Forest did not generate a probabilistic distribution similar to the original data. To address this issue, we employed an age group-wise aggregated probabilistic substitution. This approach preserved the distribution of the values across different age groups, ensuring that the overall distribution remained consistent after the substitution. Using age-specific distributions for imputation was crucial, as the likelihood of prednisone or cortisone usage may vary across different age groups due to factors like medical conditions and treatment patterns.
7. **Target Variable (KIOQ22):** This column serves as our target variable, indicating whether the respondent has been told by a health professional that they had weak or failing kidneys. It contained three different values: 1 (denoting weak or failing kidneys), 2 (denoting no kidney issues), and 9 (denoting uncertainty from the health professional). To ensure clarity in our target variable for training the dataset, we removed the 9 rows that had a value of 9 in the KIOQ22 column, thus eliminating ambiguity.

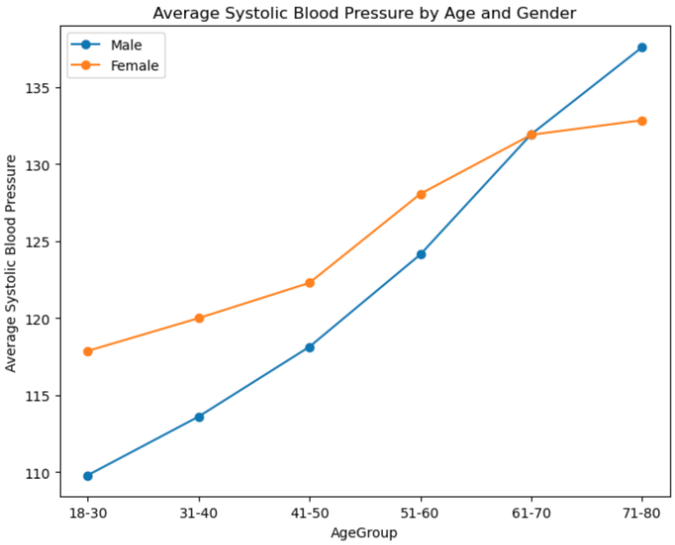
By carefully selecting appropriate imputation strategies tailored to the specific characteristics of each variable, our data cleaning process aimed to preserve the integrity of the dataset while minimizing the impact of outliers and accounting for potential variations across different groups or distributions. The chosen methods prioritized the use of robust central measures like medians, which are less susceptible to the influence of extreme values, and incorporated age-specific considerations where relevant, ensuring that the imputed values were as accurate and representative as possible. This meticulous approach to data cleaning has provided us with a reliable foundation for our subsequent predictive modeling efforts.

Exploratory Data Analysis:

In our role as data analysts for an insurance company, we conducted exploratory data analysis to understand the relationships between kidney failure and various factors such as age, gender, blood pressure, and other health parameters. To facilitate our analysis and create more meaningful visualizations, we first created an age group column.

Blood pressure and cholesterol levels are crucial indicators of overall health and can significantly impact kidney function. As an insurance company, it's essential to understand how these parameters vary with age and gender, as this information can inform the tailoring of insurance packages and pricing based on associated risks. With this in mind, we explored the data to answer several key questions, including:

1. Does systolic blood pressure depend on gender and age?



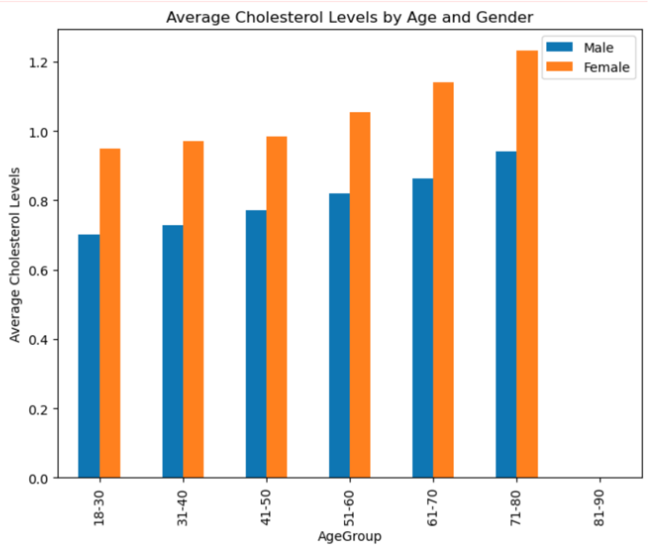
Observations:

Our analysis of systolic blood pressure patterns revealed significant age and gender-related trends. For individuals up to 70 years old, females generally exhibit higher average systolic blood pressure compared to males, suggesting a potentially higher risk for blood pressure-related health conditions, including kidney dysfunction. However, this trend reverses after age 70, with males showing slightly higher average systolic blood pressure than females. This shift indicates a change in risk factors for older adults, with elderly men potentially facing increased health risks related to high blood pressure. These findings highlight the complex interplay between age, gender, and health risks, particularly concerning blood pressure and its potential impact on kidney function.

Recommendation:

Based on these observations, we recommend that the company implement a more nuanced, age- and gender-specific approach to risk assessment and insurance package design. This could include developing separate risk models for individuals under and over 70 years of age, considering gender-specific pricing strategies, and introducing preventive care incentives focused on blood pressure management. We also suggest implementing a dynamic pricing model that adjusts premiums based on age and gender.

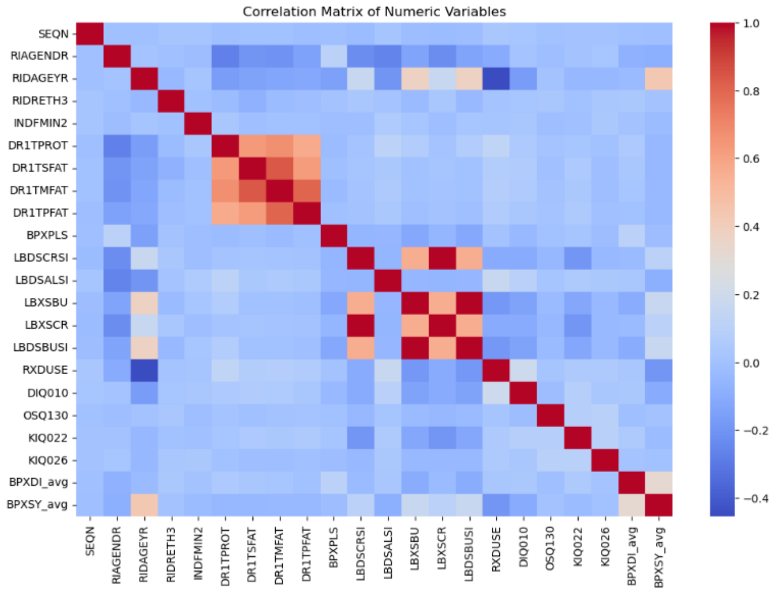
2. Does Cholesterol levels depend on gender and age of the person?



Observation: Our analysis revealed a significant gender disparity in cholesterol levels across all age groups. Females consistently exhibited higher cholesterol levels compared to males, regardless of age. This trend persisted throughout the lifespan, with the gap between genders remaining relatively stable. The elevated cholesterol levels in females suggest they may be at a higher risk for various health conditions associated with high cholesterol, including cardiovascular diseases. Given the intricate relationship between cardiovascular health and kidney function, this finding has important implications for kidney-related health risks.

Recommendation: Based on our findings, we recommend that the insurance company implement gender-specific risk assessment and policy design for kidney failure coverage. This approach should include tiered pricing structures reflecting higher cholesterol-related risks for female policyholders, alongside targeted preventive care benefits such as more frequent cholesterol screenings and nutritional counseling. However, it's important to note that these results could be specific to the North American region and could differ in other regions like Europe or Asia due to dietary, physical health, and cultural differences. Therefore, it is crucial for the company to perform region-wise analysis before launching their plans in other regions.

After doing the above analysis process, we employed visual techniques to uncover patterns and relationships between variables. We plotted the correlation matrix to unveil several intriguing observations.



Age (RIDAGEYR) exhibited moderate positive correlations with blood pressure variables and moderate negative correlations with factors such as protein intake and prescription medication use. This finding suggested that age could be an influential factor, potentially associated with changes in blood pressure, dietary habits, and medication usage, all of which can impact kidney function. Furthermore, race/ethnicity (RIDRETHS) demonstrated a moderate negative correlation with diabetes diagnosis (DIQ010), indicating a possible link between these variables and their potential impact on kidney failure risk. Notably, the blood pressure variables (BPXDI and BPXSY) displayed moderate positive correlations with age and kidney condition (KIQ022). This observation aligned with the well-established understanding that high blood pressure, or hypertension, is a significant risk factor for kidney disease, further underscoring the relevance of these variables in predicting kidney failure. Additionally, the kidney function score (KIQ026) exhibited moderate negative correlations with age and kidney condition, suggesting a potential interplay between these factors and their influence on kidney health.

To develop predictive models, two approaches were undertaken: logistic regression with backward feature selection and gradient boosting. The logistic regression model with backward feature selection aimed to identify only the most relevant variables for predicting kidney failure by iteratively removing insignificant predictors based on p-values. On the other hand, the gradient boosting model leveraged an ensemble of decision trees to capture non-linear relationships and complex interactions between variables, providing a flexible approach to model intricate patterns in the data.

Logistic Regression: Backward Selection Model

The logistic regression model with backward feature selection has identified several significant variables that can impact the prediction of kidney failure risk. These variables hold crucial implications for healthcare organizations and professionals in assessing an individual's likelihood of developing kidney disease and implementing preventive measures accordingly.

The backward feature selection was chosen for its ability to identify the most relevant predictors. Backward selection iteratively removes insignificant variables based on the Akaike Information Criterion (AIC) and p-values. This approach is particularly advantageous when dealing with many potential predictors, as it helps to eliminate redundant or irrelevant variables, improving model interpretability and mitigating overfitting. Alternative models like stepwise regression or regularization techniques could have been employed, but backward selection is a more straightforward and widely used method for variable selection in logistic regression models. Logistic regression is a type of statistical model, often used for classification and predictive analytics. This algorithm works perfectly when the Target variable is binary in nature, which is exactly what the business wants and hence we decided to go ahead and execute this model.

Gradient Boosting Model

The gradient boosting model, an ensemble learning technique that combines multiple decision trees, offers a powerful approach to capturing non-linear relationships and complex interactions between variables. While the specific variable importance scores may vary depending on the implementation, the gradient boosting model generally considers similar predictors as the logistic regression model, but with the added advantage of capturing intricate patterns and dependencies.

The model was selected for its capability to capture non-linear relationships and complex interactions between variables, which are often present in real-world data. Unlike logistic regression, which assumes a linear relationship between the predictors and the log-odds of the outcome, gradient boosting models can work with intricate patterns by combining multiple weak learners (decision trees) in an additive and iterative manner. This flexibility allows the model to adapt to complex data structures, potentially improving predictive performance. While other non-linear models like random forests or neural networks could have been considered, gradient boosting models is effective in handling classification data making them a suitable choice for this analysis.

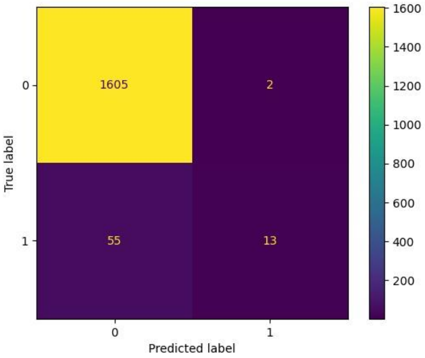
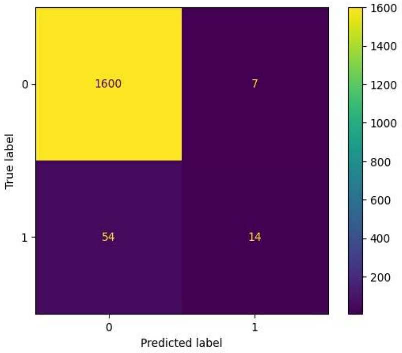
Preliminary Observations:

Gender, Age, a few Biomarkers like creatinine and albumin levels, Diabetes, Blood Pressure, prednisone and cortisone intake are the most important features which drive the Kidney Failures are selected by the Backward Selection Algorithm. Hence the data engineering team of the insurance firm needs to make sure that the data is available at least for these columns which are directly related to the Target Variable

Interpretation

Comparison of Logistic Regression and Gradient Boosting Models:

Metric	Logistic Regression	Gradient Boosting

Confusion Matrix		
Accuracy	96.60%	96.46%
Precision	86.67%	66.67%
Recall	19.12%	20.59%

Logistic Regression:
Accuracy: 96.60% - The model correctly predicted the kidney failure status for 1618 out of 1675 cases.
Precision: 86.67% - When the model predicted kidney failure, it was correct 86.67% of the time.
Recall: 19.12% - The model correctly identified only 19.12% of the actual kidney failure cases. This means that it missed 80.88% of the kidney failure cases, which is a significant risk.
False Negative Rate: 80.88% - The model failed to identify 80.88% of actual kidney failure cases, which is highly problematic as it means most kidney failure cases are not being flagged.

Gradient Boosting Model:
Accuracy: 96.46% - The model correctly predicted the kidney failure status for 1614 out of 1675 cases.
Precision: 66.67% - When the model predicted kidney failure, it was correct 66.67% of the time.
Recall: 20.59% - The model correctly identified 20.59% of the actual kidney failure cases. This means that it missed 79.41% of the kidney failure cases, which is still a significant risk.
False Negative Rate: 79.41% - The model failed to identify 79.41% of actual kidney failure cases, which is also highly problematic.

Recommendations

Based on the analysis conducted for predicting kidney failure using logistic regression and gradient boosting models, several key recommendations can be made to enhance the accuracy and effectiveness of future predictive models for kidney health assessment:

Incorporate Comprehensive Health Tests:
It is recommended to incorporate essential health tests such as HB1AC, Urine Test, and regular Blood Pressure monitoring as mandatory requirements for all policyholders. These tests provide critical insights into kidney health and can significantly improve the accuracy of risk assessment models. By making these tests mandatory, the insurance firm can ensure a more thorough evaluation of everyone's kidney health status, thereby enabling precise adjustments in insurance premiums based on actual health risks identified.

Enhance Data Collection Practices:
Improving the form and quantity of data collected is crucial for developing more robust predictive models. The current models faced limitations due to missing data points, which led to the exclusion of important features. By standardizing and enhancing data gathering processes, the insurance firm can gather comprehensive datasets that include previously missing variables. This would facilitate better analysis of trends, identification of nuanced risks related to kidney health, and informed policy adjustments aimed at optimizing risk management strategies.

Prioritize Key Demographic and Health Factors:
Age, gender, blood pressure, and cholesterol levels were identified as significant drivers of kidney health outcomes. It is recommended to prioritize these factors in risk assessment models. By categorizing policyholders into groups based on these factors (such as safe, medium, and vulnerable categories), the insurance firm can tailor insurance premiums to reflect the varying risks associated with kidney health. This personalized approach not only enhances customer satisfaction by offering fair premiums but also strengthens risk management by aligning premiums more closely with actual health risks.

Conclusion

This project leveraged the NHANES dataset to develop predictive models for kidney failure risk, employing logistic regression and gradient boosting techniques. The analysis identified key factors influencing kidney health, including age, gender, biomarkers, blood pressure, and medication usage. While both models showed high accuracy, they struggled with recall, indicating a need for further refinement. The findings highlight the importance of comprehensive health tests, improved data collection practices, and personalized risk assessment in insurance pricing strategies. Overall, this work provides a foundation for enhancing kidney disease prevention and management strategies in the health insurance industry.

Reference

- Podadera-Herreros, A., Alcala-Diaz, J. F., Gutierrez-Mariscal, F. M., Jimenez-Torres, J., Cruz-Ares, S., Arenas-de Larriva, A. P., Cardelo, M. P., Torres-Peña, J. D., Luque, R. M., Ordovas, J. M., Delgado-Lista, J., Lopez-Miranda, J., & Yubero-Serrano, E. M. (2022). Long-term consumption of a mediterranean diet or a low-fat diet on kidney function in coronary heart disease patients: The CORDIOPREV randomized controlled trial. Clinical nutrition (Edinburgh, Scotland), 41(2), 552–559. https://doi.org/10.1016/j.clnu.2021.12.041
- Whittaker, J., & Wu, K. (2022). Low-fat diets and testosterone in men: Systematic review and meta-analysis of intervention studies. arXiv:2204.00007 [q-bio.QM]. https://doi.org/10.48550/arXiv.2204.00007
- Langner, T., Östling, A., Maldonis, L., Karlsson, A., Olmo, D., Lindgren, D., Wallin, A., Lundin, L., Strand, R., Ahlström, H., & Kullberg, J. (2020). Kidney segmentation in neck-to-knee body MRI of 40,000 UK Biobank participants. arXiv:2006.06996 [q-bio.QM]. https://arxiv.org/abs/2006.06996v1
- Akinleye, A., Oremade, O., & Xu, X. (2024). Exposure to low levels of heavy metals and chronic kidney disease in the US population: A cross-sectional study. PLOS ONE, 19(4), e0288190. https://doi.org/10.1371/journal.pone.0288190
- Xia, L., Nan, B., & Li, Y. (2022). De-biased lasso for stratified Cox models with application to the national kidney transplant data. arXiv:2211.08868 [stat.ME]. https://arxiv.org/abs/2211.08868v1