Title:

BIS 687

Carrie Yao, Hang Li, Paul Zhang

Introduction

Hypertension, commonly known as high blood pressure, is a leading risk factor for cardiovascular diseases and a significant cause of premature death worldwide. Despite its prevalence, the intricacies of its management and outcomes remain complex, influenced by a myriad of clinical factors. With the advancement of healthcare technologies and the proliferation of data, there is a pressing need to harness these resources to enhance our understanding and treatment of hypertension. This project seeks to bridge this gap by utilizing the Medical Information Mart for Intensive Care III (MIMIC-III) dataset, combined with sophisticated analytical tools such as R, Python, SQL, and Tableau.

The primary objective of this research is to analyze hypertension and its clinical outcomes, such as Length of Stay (LOS) and mortality, extensively. By employing machine learning algorithms and data analysis techniques, this study aims to provide data-driven evidence support for hypertension management, improve hypertension-related clinical outcomes, optimize hospital resources, and support the development of personalized treatment plans that are more effective at the individual level.

This project is structured around three core aims: 1) Exploratory Data Analysis and Feature Engineering; 2) Predictive Analysis of Mortality and LOS; 3) Shiny App and Tableau Dashboard Development. Through this exploration, this project aims to contribute to the transformative potential of data science in healthcare, particularly in the domain of chronic disease management and hospital resources management.

Significance

Hypertension, or high blood pressure, is a pervasive public health issue in the United States, affecting nearly half of the adult population. The Centers for Disease Control and Prevention (2023) report that 48.1% of U.S. adults live with this condition. In 2021 alone, hypertension was noted as a primary or contributing cause in approximately 691,095 deaths, underscoring its deadly impact. Economically, the burden of hypertension is equally staggering, with an estimated annual cost of $131 billion from 2003 to 2014. Given the likelihood of most Americans developing hypertension during their lifetime, the necessity for early preventive measures and effective management strategies is clear and urgent (Nguyen et al.).

Our research holds profound significance in public health and machine learning application in the healthcare domain. By identifying and understanding the variables that influence LOS and mortality, healthcare providers can develop more targeted intervention strategies aimed at those most at risk. This not only enhances the precision and effectiveness of hypertension management but also significantly improves patient outcomes. Moreover, reducing the LOS in hospitals directly correlates with decreased healthcare costs and better resource allocation, which is vital for the sustainability of healthcare systems. Additionally, insights gained from this research can guide policy makers in crafting effective public health strategies and regulations to mitigate the impact of hypertension on a larger scale. Thus, the implications of this study are critical, including elevating the standard of care, optimizing healthcare expenditures, and potentially saving lives by enabling more effective management of a condition that affects a substantial portion of the population.

Innovation

The innovative aspect of our research project lies primarily in the application of multiple models that are specifically tailored to understand hypertension more effectively. This innovation is revealed through the use of cutting-edge data visualization and interactive tool development, utilizing platforms like Tableau and R Shiny to bring the findings from predictive models into a practical, usable format for clinicians and policymakers. The Tableau dashboard provides a dynamic and visually engaging interface where complex data is transformed into accessible and actionable insights. This dashboard can significantly enhance decision-making processes by offering real-time data visualization that supports the monitoring and analysis of hypertension trends and outcomes. Its intuitive layout and interactive capabilities allow users to explore data layers with ease, making it an invaluable tool for quick and informed decision-making. In parallel, the R Shiny application represents a leap forward in personalized medicine. This tool allows healthcare providers to input specific patient data and receive instant predictions regarding length of stay (LOS) and mortality risks. Such real-time predictive capabilities are crucial for the timely and effective management of hypertensive patients, enabling healthcare professionals to make informed decisions that are tailored to the individual characteristics and health status of each patient. This can lead to more personalized treatment plans that better suit the patient's specific needs and effective hospital resource management, which improves overall patient outcomes and operational efficiency.

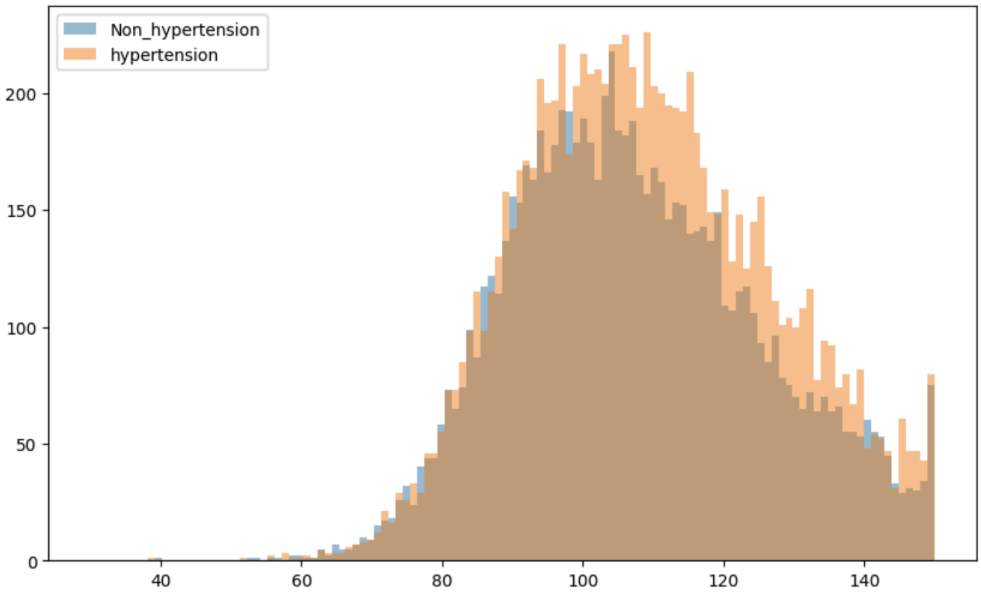
Specific Aim 1: Exploratory Data Analysis and Feature Engineering

- Method

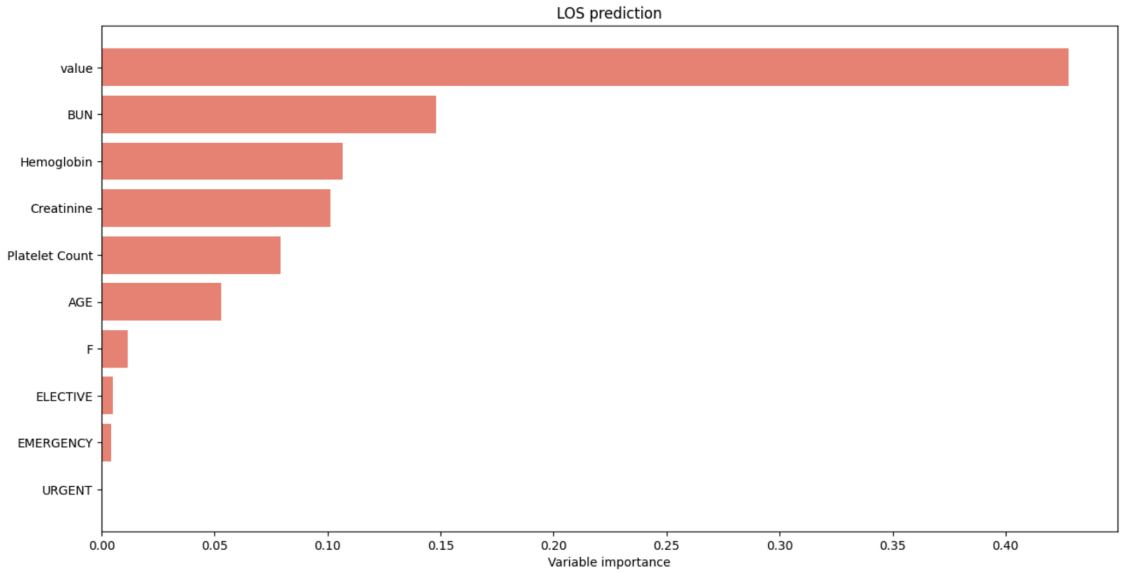
  Database Management : We upload the MIMIC database to Google Cloud Platform and create a client to link to Google Colab. In this way, we can type SQL queries to acquire data from the platform.

Feature Engineering : We select variables that are theoretically related to the likelihood of hypertension. The features include ethical features (gender, marital status,etc.), which are mostly categorical variables and lab event variables including test results like BUN, blood pressure, etc. To select variable in a less time-consuming way, we implement a random forest model and list the variable importance to select features.

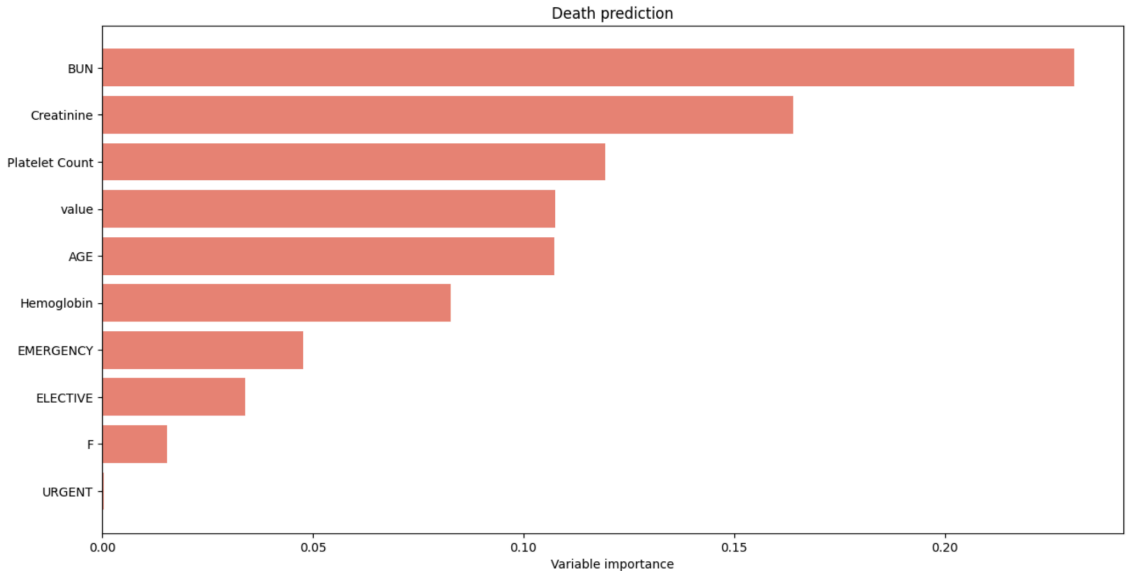
- # Results



We select the max blood pressure for both identified as hypertension and not hypertension and compare their blood pressure distribution. From the plots, we can tell that there is a significant difference in 110-120. Therefore, focusing on the length of stay and mortality rate of hypertension patients is meaningful.



The above charts describe the variable importance when predicting the length of stay. We select the blood pressure, BUN, Hemoglobin, Creatinine, Platelet Count and Age as features to predict the length of stay.



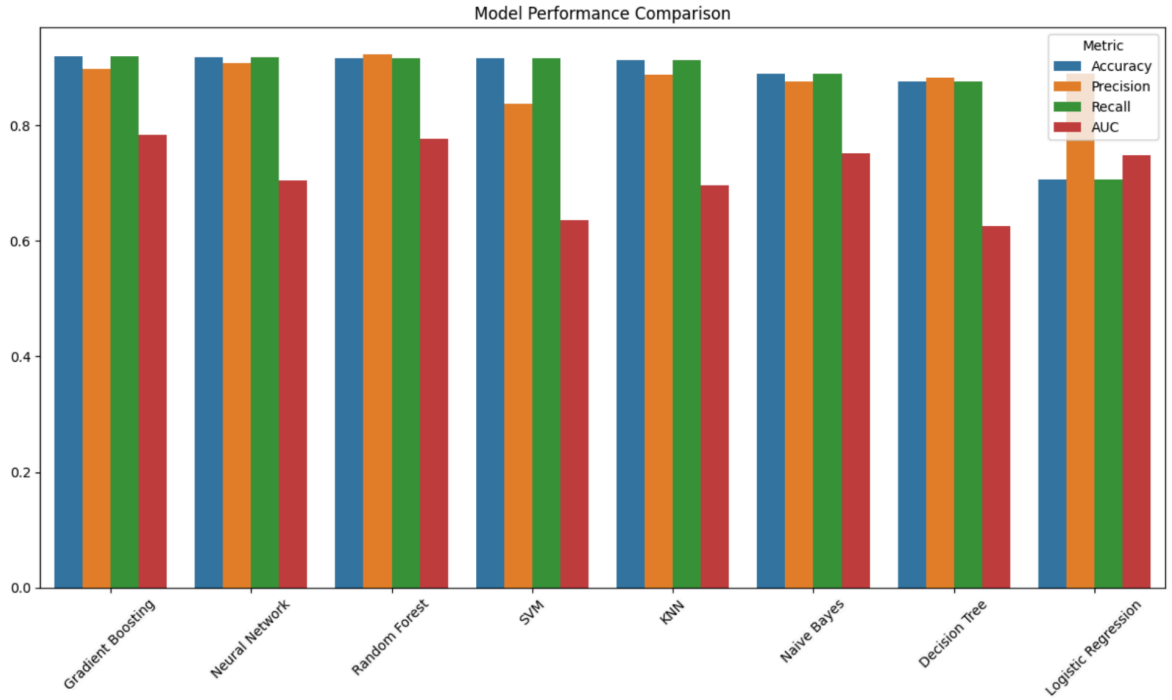
The above charts describe the variable importance when predicting the mortality rate. We select BUN, Creatinine, Platelet Count, the blood pressure, Age and Hemoglobin, as features to predict the mortality rate.

Specific Aim 2: Predictive Analysis of Mortality and LOS

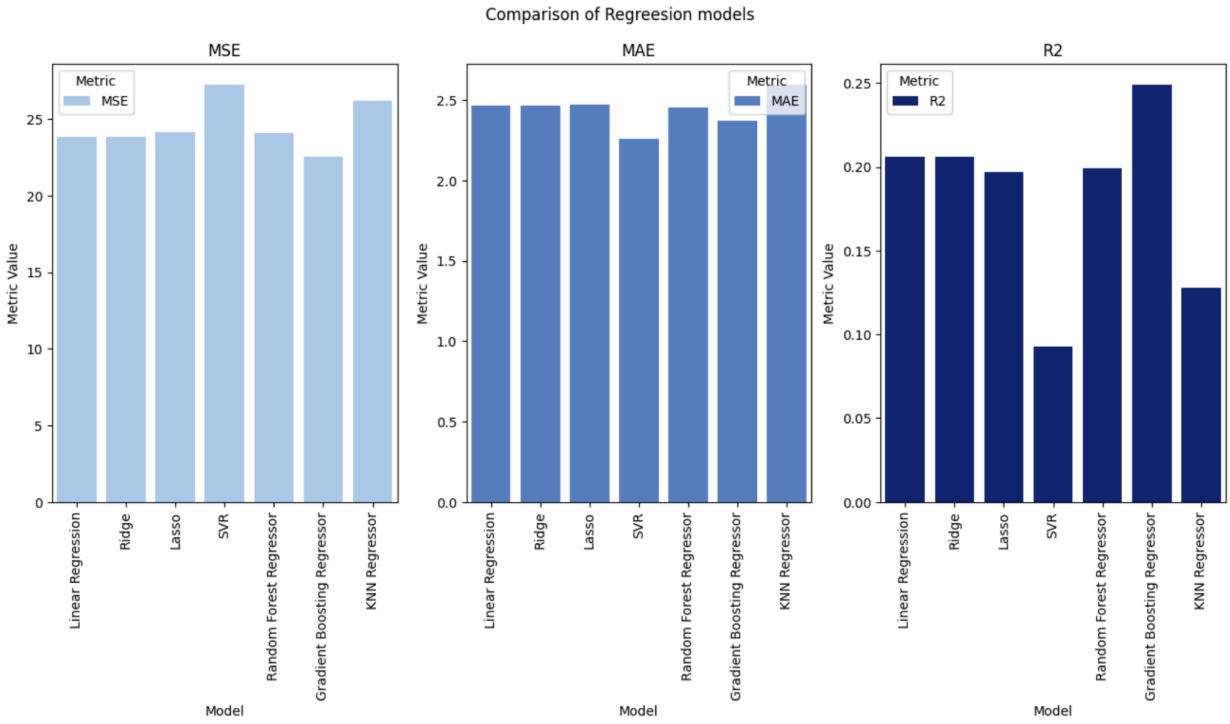
- ## Method

Model construction: Based on the features derived from part1, we establish several machine learnings respectively for two prediction aims. For predicting the mortality rate, we introduce Gradient Boosting, Neural Network, Random Forest, Support Vector Machine, K-nearest Neighbor, Native Bayes, Decision Tree and Logistic regression. For predicting length of stay, we introduce Linear regression, Ridge regression, Lasso regression, Support Vector Regressor, Random Forest Regressor, Gradient Boosting Regressor, and K-nearest Neighbor regression.
Performance evaluation: We use accuracy, prediction, recall and AUC to evaluate the performance of classification performance on mortality rate. We use mean squared error, mean absolute value and R square to evaluate the performance of the regression model on length of stay.

- Results



From the above chart, we can tell that the logistic regression model is the worst and since there is imbalance in mortality data, the accuracy of other models are all pretty high. Therefore, we mainly focus on the recall and precision and we select the gradient boosting model as the best classifier.



From the above chart we can conclude that since the R-square are all pretty low, all of the models do not perform well on this task, especially for SVR and KNN regressor. We think the reason should be that both of them are related to the spatial distribution of the feature vectors. Fortunately, we have the gradient boosting regressor that has the largest R-square and relatively low MAE and MSE and is recognized as the best regression model.

Specific Aim 3: Tableau Dashboard and Interactive App Development

- Method

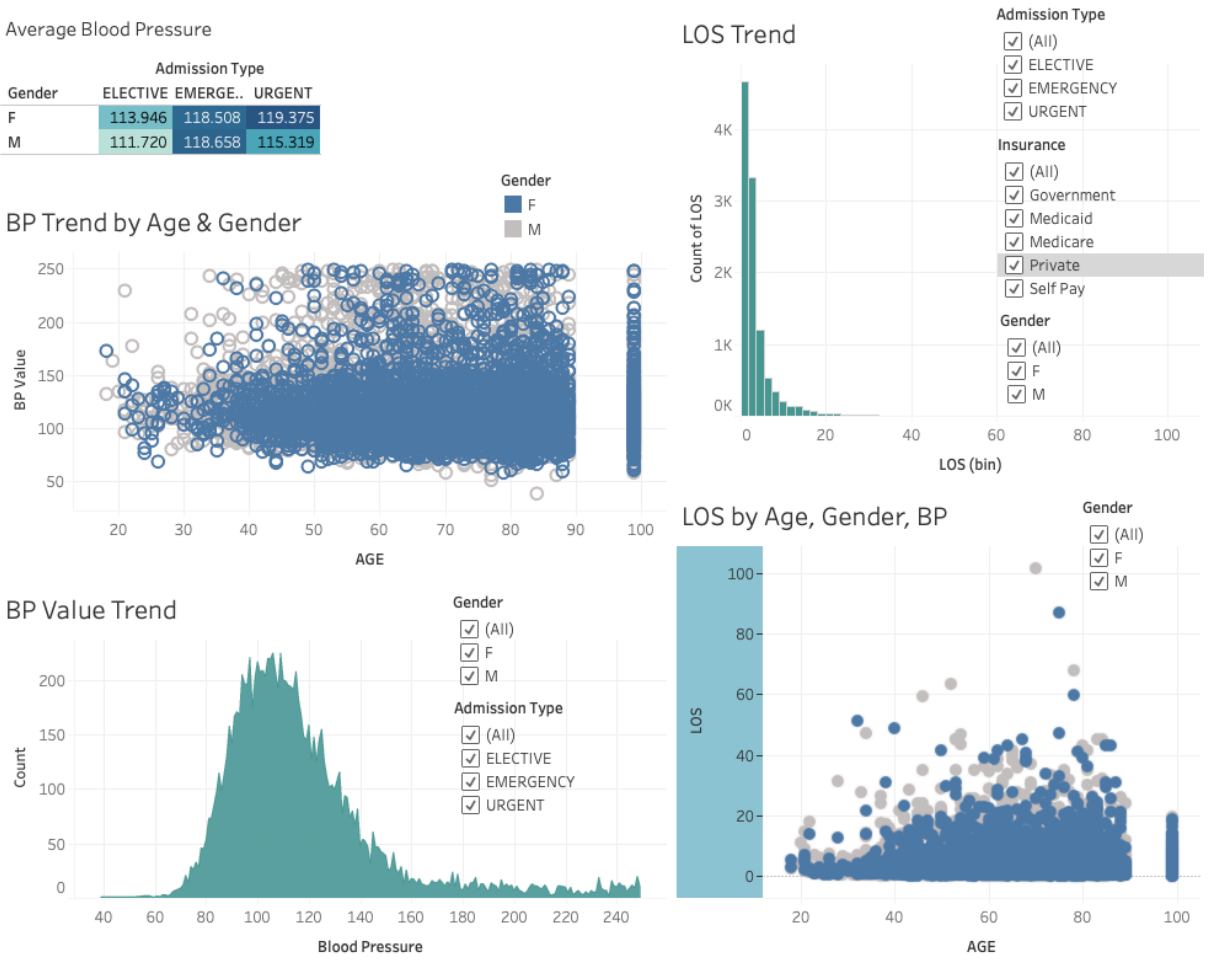
Tableau Dashboard: We used datasets extracted from the MIMIC-III database, including hypertension patients demographics dataset and hypertension patients lab results dataset. Features in these datasets include patients' age, gender, admission type, insurance, length of stay, and lab results values. We exploited these features and developed visualizations such as table, histogram and scatterplots. Using functions in Tableau, we designed an interactive dashboard that could support the monitoring and analysis of hypertension trends and outcomes.

Interactive Application: One of our initial goals was to present our predictive models through an interactive R Shiny app. Shiny apps are powerful tools that allow users to engage with complex models in a user-friendly way, making our research insights more accessible and actionable for healthcare providers. However, due to the scope of this project and our work within the Google Colab environment to connect to the MIMIC database in the cloud, we encountered some technical limitations. Running R Shiny or any other server-hosted interactive web page is not feasible within the Colab framework.

To overcome this challenge, we pivoted our approach and leveraged Jupyter widgets to simulate a streamlined version of the Shiny app experience. While not as fully featured as a dedicated Shiny app, these widgets still allow users to interact with our models and explore the impact of different patient factors on outcomes.

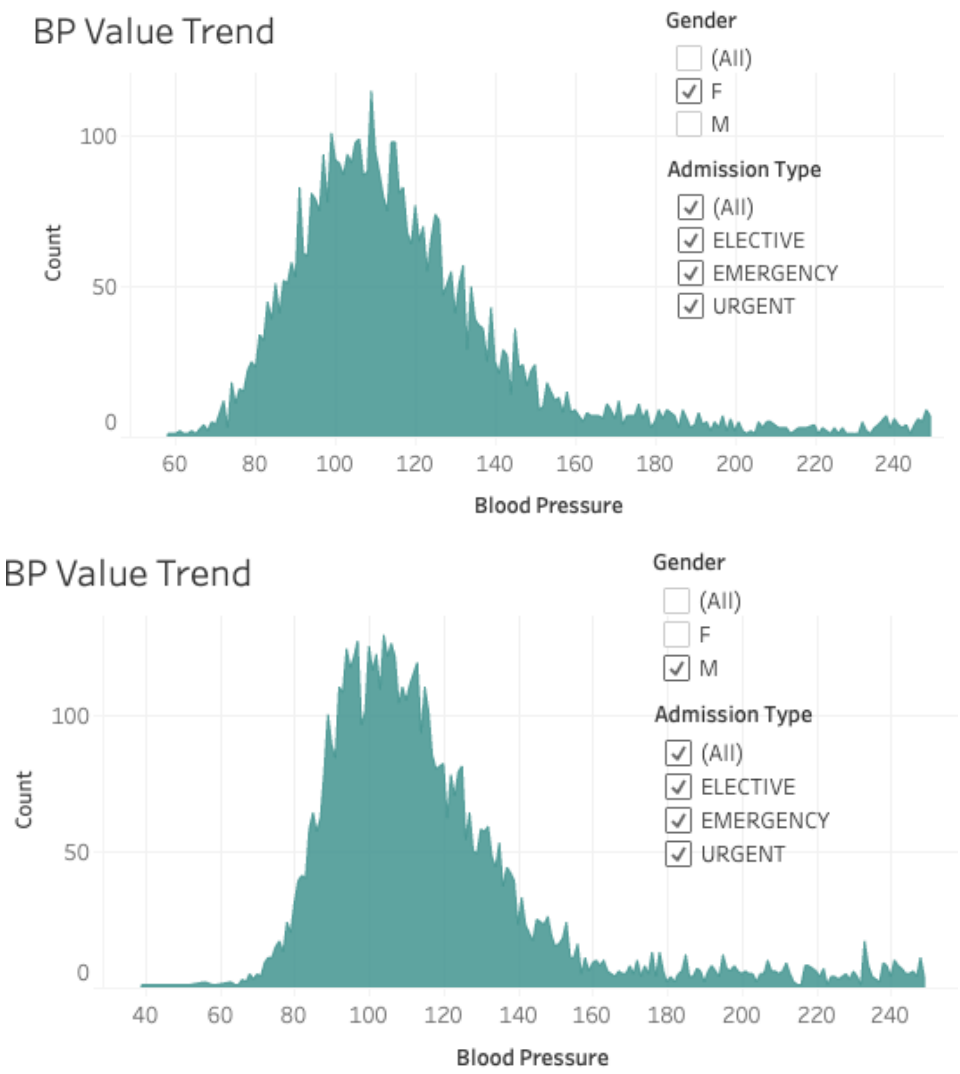
- Results

Tableau Dashboard:

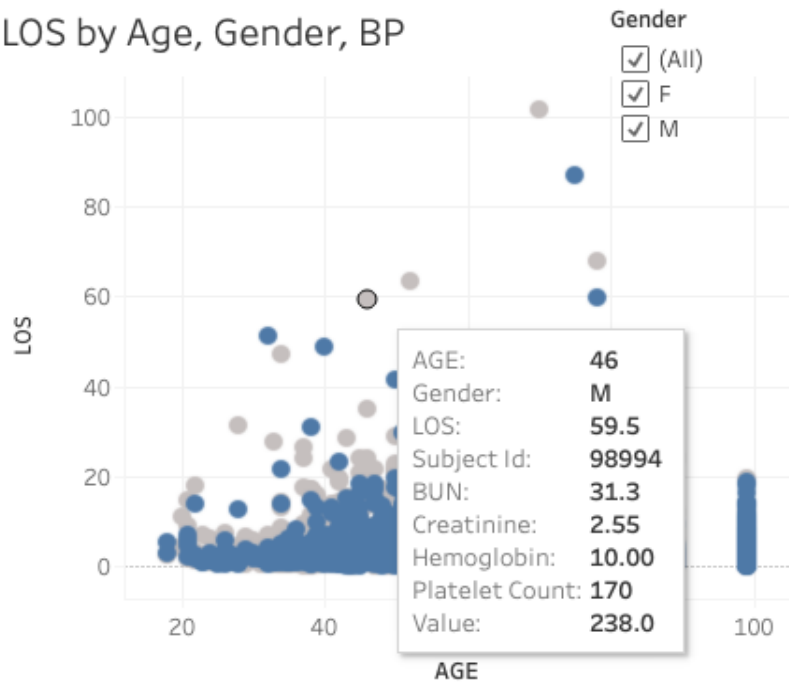


Our Tableau dashboard features interactive plots that enhance user engagement and data accessibility. This functionality not only makes it easier for users to explore and interpret complex datasets but also allows for a more detailed and immediate understanding of the correlations and trends within the data. On the top left corner of the dashboard, we've categorized average blood pressure readings by gender and admission type. This allows us to find patterns and possibly tailor interventions depending on the urgency of admission. The BP Value Trend histogram demonstrates the distribution of blood pressure values among our patients. This visual aid is particularly helpful for recognizing common ranges and outliers in blood pressure

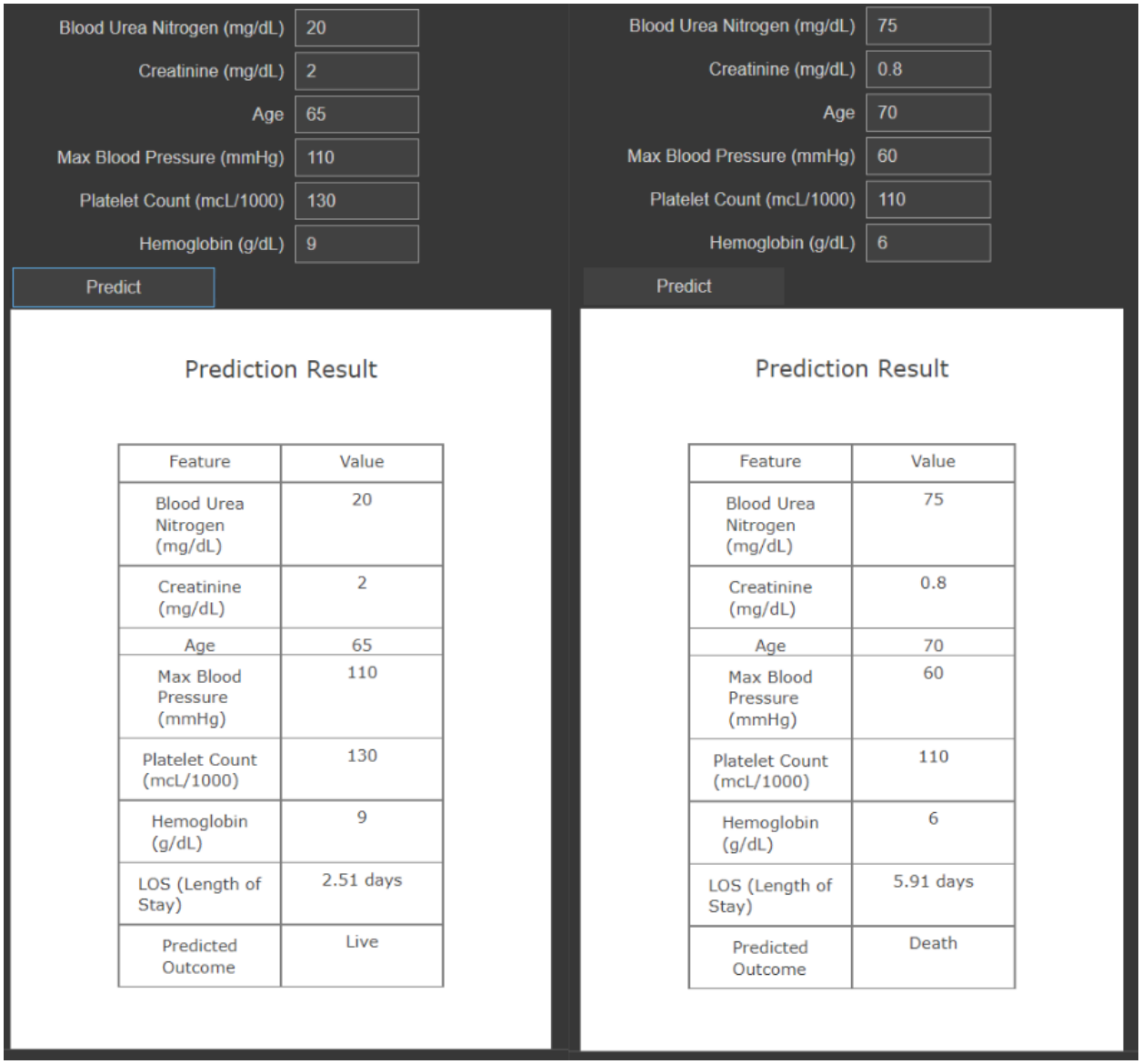
readings. With filters that enable value selection of features, we can see trends of female patients and male patients and different admission types.



To the right, the histogram shows the distribution of Length of Stay (LOS) for patients. The filters of admission type, insurance, and gender enable faceted outlook of the trend of LOS. The bottom right scatter graph gives us valuable insights into BP, LOS, lab results, age, gender of a specific patient, including lab metrics that were selected from our feature engineering process, which is crucial for healthcare planning, resource allocation, and hypertension management decision-making. Users can interact directly with the plot by hovering over or clicking on any point within the scatter plot. Upon doing so, a tooltip appears, displaying all relevant information about that specific data point of the patient.



Interactive Application: It's important to note that adapting our code for a full-fledged R Shiny app is a straightforward process. We have provided sample code on our GitHub page to demonstrate the transition. With access to a local copy of the MIMIC database, setting up a fully functional Shiny app can be accomplished in a matter of minutes. Despite the technical constraints, our use of Jupyter widgets showcases the flexibility and adaptability of our approach. We remain committed to making our predictive models accessible and useful for healthcare providers.



As we move forward, we see great potential in the development of a standalone and self-hosted Shiny app. By enabling providers to interact with our models in a dynamic, user-friendly interface, we can amplify the impact of our research and provide a valuable tool for improving hypertension management and patient care.

Conclusion & Discussion

Our project represents a significant step forward in the understanding and management of hypertension, a pervasive health concern with substantial implications for public health and healthcare resource allocation. Through rigorous exploratory data analysis, predictive modeling, and the development of interactive tools, we have shown important correlations between patient characteristics, clinical outcomes, and hypertension management strategies. By leveraging innovative approaches such as Tableau dashboards and simulated Shiny app experiences, we aim to empower healthcare providers with actionable insights for optimizing treatment plans, enhancing patient care efficiency, and ultimately improving outcomes for individuals affected by hypertension. Thank you for your attention and support.

References

Centers for Disease Control and Prevention. (2023, July 6). Facts about hypertension. Centers for Disease Control and Prevention. https://www.cdc.gov/bloodpressure/facts.htm#:~:text=In%202021%2C%20hypertension%20was%20a,deaths%20in%20the%20United%20States.&text=Nearly%20half%20of%20adults%20have,are%20taking%20medication%20for%20hypertension

Nguyen, Q., Dominguez, J., Nguyen, L., & Gullapalli, N. (2010). Hypertension management: an update. American health & drug benefits, 3(1), 47–56.