### **Exploring Inpatient Hemoglobin A1C Impact on Diabetes Care Outcomes: A Comprehensive Analysis of Electronic Health Records**

Acadia G. Berry

A Thesis Proposal

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science in Data Science

Department of Mathematical Sciences

Central Connecticut State University

New Britain, Connecticut

May, 2024

Thesis Advisor: Dr. Krishna Saha

Keywords: Diabetes mellitus, inpatient care, hemoglobin A1c, healthcare analytics, multivariable logistic regression, XGBoost, SMOTE, R

#### ACKNOWLEDGEMENTS

I would like to thank my family and friends, especially my parents Deborah and Bryan Berry, for their support as I worked towards completing this degree and thesis. Without their consistent encouragement and patience, it’s unlikely that I would have made it this far.

Next I’d like to thank Dr. Krishna Saha for agreeing to be my thesis advisor and providing excellent guidance through this entire process. Thank you also to Dr. Gurbakhshash Singh and Dr. Darius Dziuda for serving on the thesis committee. During my time in the program, I've had the privilege of being instructed by each member of the thesis committee, from whom I've gained invaluable knowledge and expertise.

Finally, I thank all my peers in this program for their active involvement and innovative contributions, which made the program truly rewarding.

#### ABSTRACT

This research investigates over 101,000 hospital records to explore the relationship between comorbidities, markers of hospital care, and early hospital readmission rates among diabetic patients. Extracted from the Cerner Health Facts Database, the dataset encompasses 10 years (1999-2008) of patient encounters across 54 United States hospitals. Leveraging multivariable logistic regression and other machine learning techniques including XGBoost, random forest, and support vector machines, the study assesses how significant variables impact early hospital readmission among this population.

Building upon previous research highlighting the association between hemoglobin A1c measurement during patient encounters and reduced readmission rates among diabetic patients, this study aims to extend these findings. Through data preparation, exploratory data analysis and statistical modeling, the findings of this research provide insights that could assist in tailoring interventions and enhanced patient care strategies. By addressing the challenges posed by heterogeneous clinical data and employing advanced analytical techniques, this study contributes to the growing body of knowledge aimed at improving healthcare outcomes and reducing healthcare costs for diabetic patients.

#### TABLE OF CONTENTS

#### ACKNOWLEDGEMENTS ............................................................................................... 3

#### ABSTRACT...................................................................................................................... 3

#### INTRODUCTION ............................................................................................................ 5

#### RESEARCH OBJECTIVES AND PURPOSE................................................................... 6

#### LITERATURE REVIEW.................................................................................................... 6

#### Background........................................................................................................... 7

#### Diabetes Treatments ............................................................................................ 8

#### Diagnostic Tests………….………………...……………………………….…………10

#### Economic Impact ………..................................................................................... 12

#### Broader Perspectives...........................................................................................13

#### INVESTIGATORY PROCESS ........................................................................................23

#### Data Acquisition....................................................................................................23

Data Preparation..................................................................................................24

Exploratory Data analysis ................................................................................... 26

Multicollinearity………………………………………………………………...……….65

Data Transformations………………………………………………………………….69

Feature Engineering……………………………………………………………...……71

Variable Selection……………………………………………………………...………80

MODELING………….......................................................................................................85

Data Balancing.................................................................................................... 94

Logistic Regression .......................................................................................... 107

Further Exploration of HbA1c......................................................................….. 115

XGBoost………….............................................................................................. 123

Random Forest.................................................................................................. 130

DISCUSSION ...............................................................................................................138

Final Model Metrics..............................................................................……….. 138

Implications for Clinical Practice……………………………………………………142

#### CONCLUSIONS & LIMITATIONS................................................................................ 145

#### GLOSSARY OF VARIABLES ..................................................................................... 149

#### REFERENCES……………….…................................................................................. 153

#### RESEARCH OBJECTIVE AND PURPOSE

This research investigates the findings of prior studies on diabetes care in US hospitals and outlines the methodology for analyzing health outcomes for diabetes patients. Variable interactions between hemoglobin A1c (HbA1c) levels and other patient factors such as demographics, comorbidities, and medication usage were assessed to provide a comprehensive understanding of early hospital readmission risks among patients with diagnosed hyperglycemia. A data set containing hospital records from 54 hospitals in the United States from 1999-2008 with over 101,000 admission cases extracted from the Cerner Health Facts database was used.

Patient readmission rate, often regarded as a benchmark of hospital performance, holds significant importance for healthcare institutions. By leveraging various machine learning algorithms, this research aims to develop classification models that can predict readmission probabilities based on a combination of clinical and non-clinical variables, achieving target values for model performance. Assessing model results provides insight into variable interactions, effects, and serves as a basis for further investigation into the impact of these variables on early hospital readmission among diabetic patients.

Identifying the impact of relevant predictors of early hospital readmission enables healthcare providers to tailor care-management strategies, enhancing patient care while minimizing unnecessary hospital visits. In turn, results may support more efficient resource allocation and reduce costs among patients and hospitals. The findings of this study can inform healthcare policies and practices related to diabetes care, ultimately advancing research in this growing area of healthcare.

**LITERATURE REVIEW**

The literature review for this analysis consists of three main areas. First, a brief explanation and background of diabetes will be provided. This section highlights the evolution and key milestones of our understanding of diabetes which have shaped the management of this complex metabolic disorder. Following this, an overview of relevant literature focused on the analysis of diabetes patient care will be presented, particularly previous findings using the data set created by Strack et. al (2014) as well as information about the source of the data. Finally, the thesis will highlight the specific data analysis and modeling techniques anticipated for identifying the key attributes contributing to patient readmission and the methods to identify patient trends.

Additionally, a glossary of the medical terms used within this data set is available in the appendix. This glossary provides the medical terminology and variables of interest used throughout this thesis. The glossary allows the readers to familiarize themselves with the intricacies of in-patient medical care, understand what type of data is collected during hospital stays and identify potential significant variables.

**Background**

Diabetes mellitus is a medical condition which affects the body's ability to process sugar in the blood. Whether it is genetic or developed over time, the unstable blood sugar levels are a result of the body’s inability to transform glucose into energy. Diabetes often requires lifelong medication to manage the symptoms, and if left untreated, it can be lethal. According to the American Diabetes Association (2024), chronic hyperglycemia can lead to various complications, including damage to the blood vessels, organs such as the kidneys, eyes and heart and the nervous system. Furthermore, individuals with diabetes are at an increased risk of developing cardiovascular diseases, stroke, and other serious health issues (American Diabetes Association, 2024).

It is unclear exactly how long diabetes has existed; however, it is known that patient diagnosis rates have risen throughout the past few decades and as of 2020, 10.5% of the US population is diabetic (Dhatariya et al., 2020). These rates are predicted to continue to rise, affecting people at younger ages with a 700% increase in type II diabetes diagnoses in people under the age of 20 by 2060 (Centers for Disease Control and Prevention, 2023). Despite these rising rates and the incurable nature of diabetes, there are preventative measures that can be taken to combat the increasing population. Diabetes education, a balanced diet, exercise, and weight management are commonly cited measures (Centers for Disease Control, 2022).

Although other rare types can occur, diabetes is most often categorized in two forms: Type 1 and Type 2. Generally, Type 1 can be characterized as more severe, diagnosed early in life and an autoimmune disease, whereas Type 2 can develop over a lifetime and is associated with insulin resistance and a dysfunctional pancreas (Matthews et al., 2008). Diabetes patients are treated using various methods including insulin therapy, oral medications, lifestyle modifications and in some cases a combination of these approaches. Type 1 diabetes typically requires insulin therapy from the time of diagnosis, as the immune system mistakenly attacks and destroys the insulin-producing beta cells in the pancreas (Owens-Collins, 2023). On the other hand, Type 2 diabetes management may begin with lifestyle changes such as diet and exercise, progressing to oral medications and eventually, insulin therapy if found necessary (American Diabetes Association, 2021).

**Treatments**

Administering insulin to diabetic patients is the most well-known solution to managing symptoms and it was only about 100 years ago in 1923 when insulin became commercially available (Bliss, 2021). Now, there are numerous methods and therapies available such as oral medications, insulin pumps and continuous glucose monitoring devices. The introduction of insulin to the medical field marked a groundbreaking milestone in diabetes management, offering a life-saving treatment for individuals with Type 1 diabetes which eventually became an essential component in the management of Type 2 diabetes. The landscape of diabetes care has evolved significantly since then, with ongoing advancements in pharmaceuticals and technology (Sugandh et al., 2023).

In addition to traditional insulin injections, oral medications play a crucial role in managing Type 2 diabetes. These medications aim to improve insulin sensitivity, reduce glucose production in the liver, and enhance glucose uptake by cells (Sugandh et al., 2023). Common classes of oral medications include metformin, sulfonylureas, and thiazolidinediones, each with its specific mechanism of action (American Diabetes Association, 2021).

Technological innovations have brought about devices like insulin pumps and continuous glucose monitoring systems. Insulin pumps provide a continuous supply of insulin, offering a more flexible and personalized approach to insulin therapy (Beck et al., 2021). Continuous glucose monitoring allows individuals to track their blood glucose levels in real-time, providing valuable data for adjusting insulin doses and making informed lifestyle decisions (Beck et al., 2021). Now, in the field of pharmacogenomics, specific genes are assessed to predict whether a person is at risk for diabetes and the effectiveness of treatments. Examining the genes associated with drug metabolism, can assist in selecting the most effective medications, optimizing dosages, and minimizing the risk of adverse reactions (Sugandh et al., 2023).

Diabetes patients require daily intervention to monitor blood-glucose levels and medications to balance these levels. Oral medications for Type 2 diabetics, insulin administration for Type 1 diabetics and even cardiovascular medication are part of medication management for millions of Americans (Mathews, 2008). It is possible to measure blood-glucose levels through self-monitoring devices, however, when in medical settings, an HbA1c test will be conducted which tracks the long-term trend in blood glucose control by the level of glycosylated hemoglobin in the blood and is appropriate for type 2 diabetics that require less severe interventions (Mathews, 2008). This health issue not only spreads far and wide, but it ranges in severity.

Despite the availability of numerous medications and methods to help manage symptoms, there is not a widely known or accepted cure. Along with daily management interventions, diabetics face high-risk outcomes, such as a 3-fold greater chance of hospitalization compared to those without diabetes and an increased risk of complications and mortality. These outcomes further highlight the need for effective, proper, timely diagnosis and intervention strategies (Dhatariya et. al, 2020).

**Diagnostic Tests**

Serving as the cornerstone for diagnosing type 2 diabetes for nearly two decades, and initially proposed by the National Diabetes Data Group in the United States and later endorsed by the World Health Organization Expert Committee on Diabetes Mellitus, the focus on diagnosing diabetes has long been through assessing blood-glucose levels either in the fasting state or after conducting an oral glucose tolerance test (OGTT) (Florkowski, 2013). However, in 1997, there was a shift from a focus on diagnostic test results capturing glucose levels over only a single point in time, to tests providing broader clinical outcomes, such as HbA1c tests (Florkowski, 2013).

Unlike fasting glucose measurements or glucose levels during an OGTT, HbA1c test results reflect a person’s average blood sugar levels over the past 2-3 months. HbA1c tests not only provide a wider range of information but require less time and resources in comparison to previous blood glucose testing options. This shift towards utilizing HbA1c as a diagnostic tool parallels the growing understanding of diabetes as a condition with far-reaching implications, emphasizing the importance of addressing long-term complications associated with the disease (Florkowski, 2013).

Florkowski (2013) highlights several advantages to utilizing the HbA1c test as a diagnostic tool for diabetes, such as its ability to reflect chronic glycemia, convenience, validity, stability, and precision. HbA1c levels can reflect chronic glycemic disfunction through highlighting glycemic control over the 120-day lifespan of red blood cells, where 50% of HbA1c levels represent the month prior to testing and 25% in the month before that (Yasuhiro et al., 1995). This measurement makes it particularly relevant for diagnosing a disease characterized by a gradual development of complications.

Researchers also argue HbA1c is more convenient and accessible as it eliminates the need for fasting and requires only a single blood sample compared to the OGTT test which requires 3 days of strict diet and test time of 2 hours. This may encourage more individuals to undergo testing, potentially improving diabetes detection rates, particularly among undiagnosed cases (Florkowski, 2013). It was also argued that HbA1c tests demonstrate remarkable stability and precision. HbA1c specimens are resilient after collection, in contrast to plasma glucose levels which require post-collection care some labs do not fulfill (Bruns et al., 2009). Lastly, HbA1c results are correlated with clinical outcomes, notably in the context of microvascular complications like diabetic retinopathy (Colagiuri et al., 2011).

Florkowski (2013) compared studies from different health organizations and found slight inconsistencies in the way HbA1c scores are interpreted. Both the World Health Organization (WHO) and the American Diabetes Association (ADA) recommends using HbA1c at a cutoff of ≥ 48 mmol/mol (≥ 6.5%) for diagnosing diabetes (World Health Organization, 2006 & Diabetes Care, 1997). The ADA recommends individuals with HbA1c levels of 39–46 mmol/mol (5.7–6.4%) be considered at increased risk for diabetes and cardiovascular disease (Diabetes Care, 1997).

The Australian Diabetes Society recommends HbA1c level of ≥ 48 mmol/mol (≥ 6.5%) as the cut-off point for diagnosing diabetes but an HbA1c level of < 48 mmol/mol (< 6.5%) does not negate a diagnosis. Furthermore, the New Zealand Society for the Study of Diabetes recommends a diagnostic cut-off rounded up to ≥ 50 mmol/mol (≥ 6.7%). Efforts have been made to standardize HbA1c measurements. Florkowski (2013) highlights the need for consistent interpretations and the possibility of using rule-out and rule-in criteria where, for example, HbA1c levels ≤ 37 mmol/mol (≤ 5.5%) can be used to rule out diabetes, while levels ≥ 53 mmol/mol (≥ 7.0%) can be used to rule in diabetes. It can be concluded, however, that the target HbA1c for most people with diabetes is less than 7% (Cooke, 2010).

**Economic Impact**

In addition to the day-to-day interventions and higher rates of morbidity which diabetics face, the rising prevalence of diabetes presents an economic burden. Diabetes and associated outcomes made up more than 1 in 5 health dollars or a total of $245 billion in 2012 (Schnell et al. 2016). This substantial financial impact includes direct medical costs, such as hospital and physician services, as well as indirect costs associated with productivity losses and disability (Schnell et al., 2016). The economic impact of diabetes reaches far outside the hospital setting.

Furthermore, a longitudinal study examining American healthcare costs from 1996 to 2013 found that in 2013, more healthcare dollars were spent on diagnosed diabetes than on any other medical condition (Dieleman et al., 2016). The rising costs associated with diabetes care are not only attributed to the direct medical expenses but also to the indirect costs arising from the complications and the comorbidities associated with the disease. These complications often lead to increased hospitalizations, more frequent outpatient visits and higher demands for medical services, contributing to the economic burden (Dieleman et al., 2016).

Moreover, the economic impact of diabetes extends beyond healthcare spending. Researchers found that of the $412.9 billion spent on diabetes health care in 2022, $106.3 billion were indirect costs (Parker et al., 2024). The major contributors to the indirect costs of diabetes are reduced employment due to disability ($28.3 billion), presenteeism ($35.8 billion) and lost productivity due to premature deaths ($32.4 billion) (Parker et al., 2024). Absenteeism is also a contributing factor, with diabetics having an average of 1.9 days off from work (Parker et al., 2024).

**Broader Perspectives**

Connecting the history of diabetes to modern healthcare practices, the previous discussion has underscored the multifaceted nature of diabetes management. From the foundational understanding of hyperglycemia to the advancements in treatment methods (insulin therapy, innovative technologies, genomics) the complexities of diabetes care are apparent. Transitioning from these broader perspectives, investigations into patient readmission provide a crucial bridge to understanding gaps in current healthcare practices.

Building upon the findings of Strack et al. (2014) this project explores the implications of HbA1c measurement in influencing hospital readmission rates among patients with diagnosed hyperglycemia. The findings of Strack et al. (2014) were replicated and validated, confirming the association between HbA1c measurement and reduced readmission rates among patients with a primary diagnosis of diabetes. There was also an investigation of other potential variable interactions such as patient demographics, medication usage and derived features like patient comorbidities and healthcare utilization in relation to the measurement of HbA1c levels.

In addition to assessing these variable relationships, a series of classification models were built to predict patient readmission. One limitation of previous research lies in the absence of detailed information regarding the performance of the models employed by Strack et. al (2014). This analysis aims to address this gap by evaluating the performance metrics not only of logistic regression models but of other machine learning methods such as decision trees, random forests, and gradient-boosting methods to determine the most effective model for predicting hospital readmission.

Researchers uncovered a notable underutilization of HbA1c tests during hospitalizations (18.4%) and revealed a correlation between these tests and reduced readmission rates, particularly for patients with a primary diagnosis of diabetes (Strack et al., 2014). They found that among individuals with a primary diagnosis of diabetes, HbA1c measurement alone, regardless of test result, was associated with lower readmissions rate compared to patients who did not receive HbA1c testing (Strack et al., 2014). Patients who did not receive HbA1c testing were also less likely to have their medication changed. These results suggest that healthcare providers may be less responsive to adjusting medications in the absence of this specific diagnostic information.

The health data, ranging 10 years, was extracted by Strack et al. (2014) using patient records from the Cerner Health Facts database. The Cerner Health Facts database, run through the University of Tennessee Health Science Center (UTHSC), contains medical records of over 50 million patients. The UTHSC Center for Biomedical Informatics, which uses this database, has a strong focus on applying cutting edge techniques for guiding the future of health. Current projects include using machine learning and neural networks on streaming physiologic data to predict sepsis or developing new machine learning methods to predict cardiac arrhythmias from normal electrocardiographic tracings (<https://www.uthsc.edu/cbmi/about/index.php>). This data can be accessed by completing the Cerner Health Facts Request Form.

The initial data set was collected by researchers that satisfied the following criteria (Strack et al., 2014, p.2):

1. It is an inpatient encounter (a hospital admission).
2. It is a diabetic encounter, an admission where any kind of diabetes was entered into the system as a diagnosis.
3. The length of the stay was at least 1 day and at most 14 days.
4. Laboratory tests were performed during the encounter.
5. Medications were administered during the encounter.

An encounter refers to an individual patient's visit to a hospital during which diabetes was recorded as a preexisting medical condition. A total of 101,766 encounters based on these criteria were identified, resulting in a dataset with 55 attributes that describe diabetic encounters, patient demographics, diagnoses, medications, and insurance information (Strack et al., 2014).

Researchers cleaned and manipulated the data to ensure the model assumptions were met (Strack et al., 2014). To reduce noise and fulfill the regression assumption of independence among observations, the records were limited to one (the first) encounter per patient. Features with large amounts of missing data were removed, such as weight (97%) and payer code (40%) and medical specialty (47%) (Strack et al., 2014). Control variables were selected to account for patient demographic and illness severity. A feature to represent readmission was created with two values: readmitted if the patient returned within 30 days of discharge and otherwise covering both readmission after 30 days and no readmission (Strack et al., 2014).

Significant covariate relationships were found between different combinations of variables. Through assessing results of a sensitivity analysis, researchers concluded discharge disposition is significantly correlated with time in the hospital, medical specialty, age, and primary diagnosis, as there were significant changes in beta-coefficients when discharge disposition was removed from the model (Strack et al., 2014). Researchers also note significant pairwise correlations as: race (𝑃 < 0.001), medical specialty of the admitting physician (𝑃 = 0.001), primary diagnosis (𝑃 = 0.005), and time in hospital (𝑃 < 0.001); the specialty of the admitting physician with time in hospital (𝑃 = 0.001) and *age* (𝑃 < 0.001); and the primary diagnosis with time in the hospital (𝑃 < 0.001) and HbA1c (𝑃 = 0.004) (Strack et al., 2014).

Strack et al. (2014) built 4 different multivariable logistic regression models:

1. Logistic regression model with all variables but HbA1c (core model)
2. Core model with HbA1c
3. Core model with significant pairwise interactions to the core model
4. Core model with significant pairwise interactions and HbA1c

The final model included the variables discharge, race, admission, medical specialty, time in hospital, age, primary diagnosis and HbA1c. HbA1c measurement was found to be significantly influenced by the primary diagnosis, with diabetes as one of the secondary diagnoses. After adjusting for covariates, researchers found the readmission profiles of patients with a primary diagnosis of diabetes mellitus differed significantly from those with primary diagnoses of circulatory diseases and approached significance for patients with primary respiratory disease diagnoses (Strack et al., 2014). Researchers compared the predicted readmission rates for these three conditions, which account for 52.4% of all records. These predictions were calculated as the average time spent in the hospital and at reference levels for other covariates.

These findings highlight the underutilization of HbA1c testing in inpatient settings with potential implications for patient outcomes (Strack et al., 2014). Authors recognize various limitations to the study, such as issues relating to data quality, the inability to establish causality, and questions of external validity. They clarified their focus on pattern identification rather than exploring the reasons behind clinical decisions and highlighted potential influences of evolving diabetes care practices on their results. In all, results call for further investigation into strategies addressing glucose control during inpatient care for individuals with diabetes.

**Modeling Approaches and Frameworks**

Researchers argue that while conventional statistical approaches have traditionally dominated the field of health informatics, the tide is shifting towards the incorporation of machine learning techniques (Artetexe et al., 2018). In a comprehensive systematic review of 77 journal articles about studies in which patient readmission was modeled, researchers found the number of yearly publications from 2012-2017 that used machine learning techniques (tree-based models, neural networks and support vector machines) increased from 0 to 38%, highlighting the emergence of machine learning as a promising avenue for improving model accuracy and the evolving landscape of readmission risk prediction (Artetexe et al., 2018).

Other researchers highlight the trade-off in incorporating ML techniques into medical settings (Ben-Assuli et al., 2018). They say that although machine learning methods (boosted decision trees, random forests, and neural networks) are more robust and produce more accurate predictions, they lack interpretability. Alternatively, classical statistical methods (logistic regression, Naive-Bayes, and single decision trees) are simple and interpretable, but not nearly as accurate (Ben-Assuli et al., 2018). This poses issues in a clinical setting, where the clarity of model interpretation is crucial.

Among the 77 studies which aimed to model patient readmission, model performance varied depending on the target population and readmission threshold. Short-term readmission thresholds (7 days or less) generally achieved better results in terms of AUC (Artetexe et al., 2018). Considering the dataset for this project contains *Readmission* categorized by either >30 Days, <30 Days and No readmission, insights may arise from assessing if models can better predict readmission for patients which returned to the hospital less than 30 days from their initial encounter vs those who were readmitted more than 30 days from their encounter.

The content analysis article also contained interesting conclusions about the methodology for modeling patient readmission. For example, for the studies that did use ML techniques, only 30% of them discussed class imbalances, which is interesting considering one of the advantages to ML techniques like random forests, is their ability to manage imbalanced data (Artetexe et al., 2018). The common oversight of managing class imbalances will be addressed in this project, particularly for the variables *A1Cresult* and *Readmitted* (see Figure 2).

Of the 77 studies, researchers found that 54% of studies modeling hospital readmission employed random forest (RF) models (Artetexe et al., 2018). Researchers also found RF successful when modeling hospital readmission on a set of patient records from the General Hospital of Komotini (Michailidis et al., 2022). Using 11,172 records with independent variables covering administrative, medical-clinical, and operational aspects, researchers created four different models: support vector machines with linear kernels, support vector machines with RBF kernels, balanced random forests and weighted random forests to predict readmissions. The balanced random forest model outperformed all others, achieving a sensitivity of 0.70 and an AUC value of 0.78 (Michailidis et al., 2022). Despite the model's effectiveness in predicting patient readmission for a general population rather than specifically for a diabetic population, the random forest algorithm still outperformed other methods and was retained as a contending model in the proposed study.

In another study, researchers used 29,702 records of adult patients who were admitted to Ghent University Hospital in 2016 (Deschepper et al., 2019). The goal of this research was to identify the most well-performing model for predicting readmission with respect to hierarchy within the ICD pathology data, which was also investigated to determine the level of detail of ICD codes necessary for the most accurate predictions. Various predictive models were evaluated, including logistic regression, penalized logistic regression, gradient boosting, and random forests.

Interestingly, researchers found having only 3-digit ICD codes versus the more detailed 5-digit ID codes produced better predictions in the modeling process. These findings suggest that across all models tested, random forests consistently outperform other methods, such as yielding a notable improvement of 7% when compared to logistic regression (Deschepper et al., 2019). This research not only suggests that an RF model may be useful in predicting readmission but also backs the logic of Strack et al. (2014) who shortened the ICD codes to 3-digits in data preprocessing steps.

Researchers suggest the high performance of RF models is due to their ability to manage imbalanced data (Deschepper et al., 2019 & Michailidis et al., 2022). In healthcare datasets, the occurrence of the positive class (readmissions) is often lower than the negative class (non-readmissions), making it challenging for some algorithms to predict the minority class (Larose & Larose, 2015). This research demonstrates the effectiveness of using a balanced RF model to overcome some of the complexities of medical data, like imbalanced distributions.

Gradient boosting machine (GBM) models will also be employed in this analysis. In a study of hospital readmission data from patients in Alberta, Canada, researchers used a population-based cohort study design with linked administrative observational data to train prediction models for 30-day all-cause hospital readmissions (Davis et al., 2022). GBM models were utilized and significantly outperformed the conventional Logistic Regression (LR) model. The GBM model, using decision trees as base learners, demonstrated a substantial improvement in Area Under the Curve (AUC) on the test set, highlighting an increase in prediction accuracy of 17% at 0.83 compared to the LR model, with an AUC of 0.66. Given the success of GBM modeling in a comparable scenario, this modeling technique will also be incorporated into the project.

In another study, researchers aimed to model readmission risk to healthcare facilities for patients with Chronic Obstructive Pulmonary Disease and asthma (Demir, 2014). The study defines numerous factors related to patients' medical history, comorbidities, and healthcare service utilization as well as over 100 additional variables derived by researchers about patient medical service history. The predictive models (logistic regression, classification trees, Generalized Additive Models (GAMs), and Multivariate Adaptive Regression Splines (MARS)) were evaluated using performance metrics such as the area under the Receiver Operating Characteristic (ROC) curve, sensitivity, specificity, Brier's score, and generalized R2N index (Demir, 2014).

Results show comparable performance between logistic regression and regression trees, with mean ROC curve areas of 0.924 and 0.928, respectively (Demir, 2014). This comparable performance suggests a CART model would be useful to incorporate into the selection of models for the project at hand. Despite the success of applying machine learning techniques in similar readmission risk scenarios, complex models like GAMs and MARS had lower predictive accuracy. Researchers suggest this is due to potential overfitting (Demir, 2014). This study highlights the importance of incorporating both classical statistical methods and newer, more complex machine learning techniques.

#### **INVESTIGATORY PROCESS**

**Data Acquisition**

The data for this analysis originates from the Cerner Health Facts database. The data set was extracted and compiled by Strack et al. (2014) and can be accessed via the University California Irvine (UCI) Machine Learning Repository. The UCI Machine Learning Repository is an archive that contains donated data sets, databases and data generators, openly accessible to the public (Kelly et al., 2023). This dataset is licensed under the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/legalcode) (CC BY 4.0) license. Access to this data is permitted under the condition that proper citations are provided. The URL to access the data is: (<https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008>).

The methodology driving data analysis in this project follows the Cross-Industry Standard Process for Data Mining (CRISP–DM), created by analysts working for Daimler-Chrysler, SPSS, and NCR (Chapman, Clinton & Kerber et al., 2000). This methodology includes the six phases: Research Understanding Phase, Data Understanding Phase, Data Preparation Phase, Modeling Phase, Evaluation Phase and Deployment phase. The CRIS*P-*DM method is iterative and adaptive, meaning each of the phases can be returned to in different steps of the analysis, however, the structure of this cycle remains the same (Larose & Larose, 2015). These phases are incorporated into various steps of the analysis.

**Data Preparation**

A screenshot of a computer

Description automatically generated

*Figure 1: Initial look at the data*

The data was downloaded as a single comma separated value (.csv) files and loaded into R Studio. Accompanying documentation provided by Strack et al. (2014) containing information about ICD-9 codes, admission source ID, admission type ID, discharge disposition, and medical specialty was also downloaded. This documentation serves to map each numeric value to its corresponding categorical label among these variables. The data contains one record per patient encounter, with the possibility of multiple encounters per patient, trackable by patient ID field. There are 50 available fields with 101,766 total records. Upon reviewing the summary of unprocessed dataset, several notable insights emerge.

Patients span a broad age range, with a concentration between 40 and 80 years old, suggesting a demographic skew towards older adults. Clinical features, including hospitalization duration, medical procedures, medications, and diagnoses, provide valuable insights into patient healthcare utilization and medical conditions. Medication usage data reveal diverse prescription patterns across different medication categories. Moreover, diabetes-related features, such as glucose serum levels, HbA1c results, and insulin, underscore the dataset's focus on diabetes management and treatment. Finally, the target variable, *readmitted* stratifies patients based on their readmission status within 30 days, after 30 days, or no readmission.

All categorical variables exhibited a proportion of missing values. Notably *payer code* and *weight* had the largest missing value rates of 39.55% and 96.85%, respectively. To address missing values in the categorical variables, a category was designated to signify absence, thus maximizing the retention of available data. Missingness itself can be indicative of underlying factors that may influence readmission risk, stemming from systemic issues such as disparities in data collection practices, differential access to healthcare services, or patient reluctance to disclose sensitive information. Recognizing and accounting for missing data as a separate category enables these potential biases to be explored.

Further data preprocessing steps were undertaken to ensure the data set is representative of the population we are interested in. To avoid potential biases that might arise from multiple observations per patient, only the first encounter per patient was retained in the dataset. This reduced the number of encounters from 101,766 to 71,514. To prevent undue influence on analysis outcomes, encounters with discharge dispositions indicating death or hospice care were excluded from the dataset, further reducing the data set size to 69,969. Including encounters that resulted in death could introduce inconsistencies in the dataset, as readmission is not applicable to patients who have passed away. Newborn-related encounters and patients with *age* 0-10 were also excluded to ensure the focus remained on adult patient populations, eliminating an additional 468 records with a set size of 69,501. The final data set includes 69,501 records of unique patient encounters.

**Exploratory Data Analysis**

The following section entails a review of steps taken as part of exploratory data analysis (EDA), such as assessing variable summaries, distributions, and the unique attributes of contending predictor variables with the cleaned data. As part of EDA, some variables underwent transformations (see page 69). Comparative analyses were conducted between the variables’ original form in the unprocessed data and the transformed variable utilized in the model, highlighting the rationale behind them. Overall, the steps taken in EDA facilitate a more comprehensive understanding of the dataset's intricacies and serves as a guide for subsequent modeling steps.

***Readmitted***

In the original dataset, *readmitted* is a categorical variable comprising of three levels: *<30, >30,* and *NO*. These categories denote whether a patient was readmitted within less than 30 days after the initial encounter, more than 30 days after the initial encounter, or not readmitted at all, respectively. The 30-day hospital readmission marker serves as an industry standard. Different partitioning methods of *readmitted* were tested to determine the most informative split for modeling. See *Figure 1* for a bar chart of *readmitted* from the original data set and *Table 1* for a table of counts.

A graph with different colored squares

Description automatically generated

*Figure 2: Bar chart of Readmitted*

|  |  |  |  |
| --- | --- | --- | --- |
| *Readmitted* | | | |
| <30 | >30 | NO | Total |
| 11,354 | 35,519 | 54,727 | 101,600 |
| 11.18% | 34.96% | 53.86% | 100% |

*Table 1: Total counts and proportions of Readmitted*

To identify the optimal classification split for predicting *readmitted* we evaluated the performance of two logistic regression models. These models were constructed based on two subsets of the data, distinguished by different definitions of positive occurrences (*y* = 1) and negative occurrences (*y* = 0) for *readmitted.* One subset categorized occurrences as positive when *readmitted* was less than 30 days (<30) and negative when *readmitted* was greater than 30 days or *NO.* The other subset excluded occurrences where *readmitted* was greater than 30 days entirely, defining positive occurrences as less than 30 days and negative occurrences as *NO*. These models used predictor variables, age*, gender, race, discharge disposition, admission source ID, admission type ID, HbA1c result, time in hospital,* and *medication change.* Due to the iterative nature of the data science methodology, the model used to evaluate the impact of this variable partition, and subsequent sensitivity analyses, is based on models developed in subsequent steps, following initial EDA.

The *>30* marker indicates readmission occurring more than 30 days but less than one year after the initial encounter. Due to the uncertainty regarding the timeline beyond 30 days, including these recordscould potentially obscure the distinctions between the readmitted groups. In the logistic regression model including records of patients who were readmitted more than 30 days after their initial encounter, model accuracy was 59.1%, with a specificity of 59.3% and a sensitivity of 57.8%. Conversely, in the model excluding records of patients readmitted more than 30 days after their initial encounter, the accuracy improved to 62.8%, showing higher specificity (63.9%), and lower sensitivity (51.97%).

These results show a model excluding records of patients who were readmitted more than 30 days later, increases model sensitivity by 5.85%. This may be because this subset encapsulates records with the most distinct information from each other, making the features that lead to these outcomes more apparent. Ultimately, it was concluded that records with levels *<30* and *NO* would form the binary split of the response variable in the final model. Figure 2 illustrates the distribution of the binary variable, readmission, in the final data set, with *Table 2* providing a variable summary. Removing records of patients who were readmitted more than 30 days after their encounter further reduced the data set to 43,508 records.

A bar chart with a bar and a bar chart

Description automatically generated

*Figure 3: Distribution of readmitted in final data set*

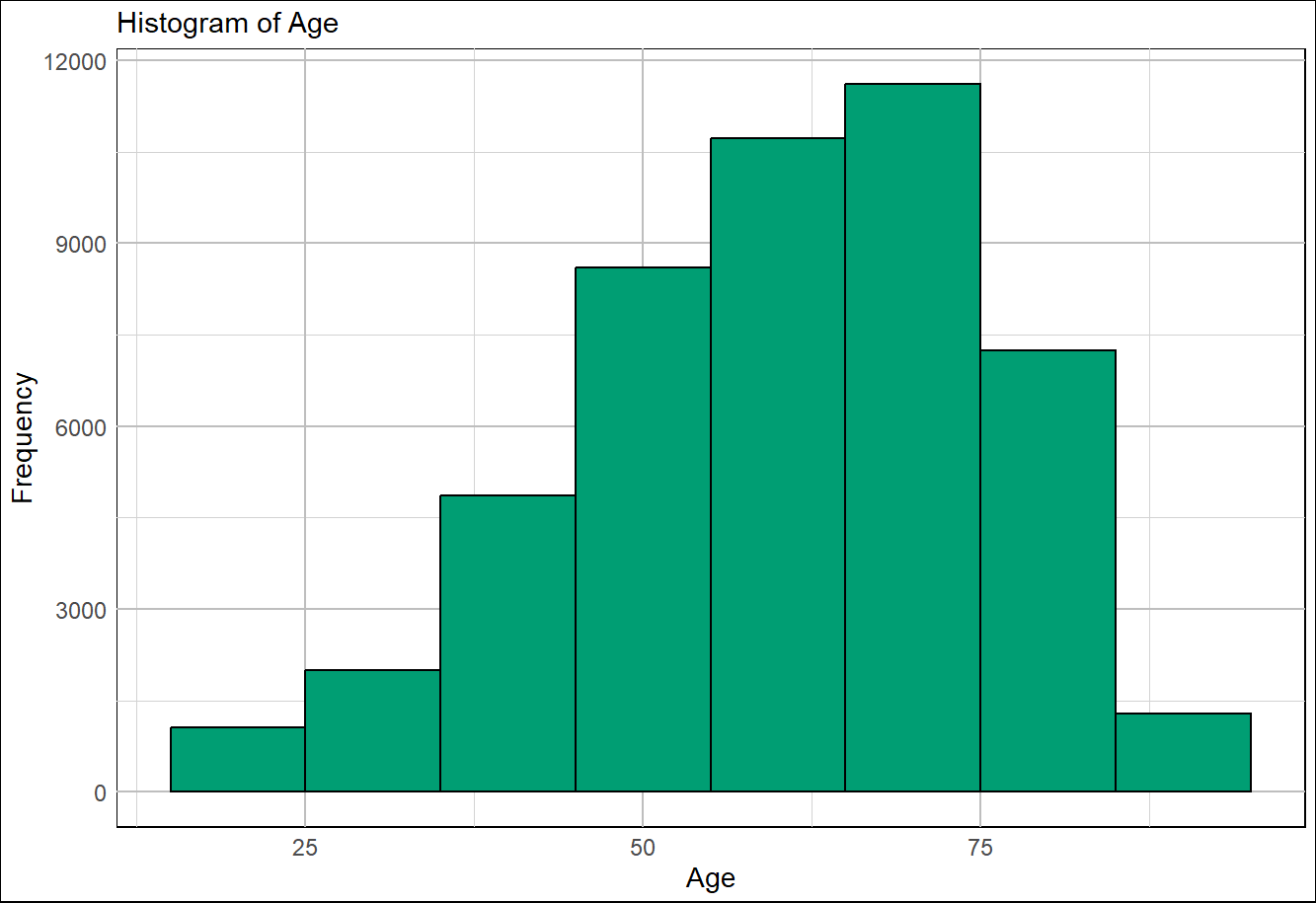
|  |  |
| --- | --- |
| *Readmitted* | |
| <30 | NO |
| 5,736 | 37,772 |
| 13.184% | 86.816% |

*Table 2: Counts and Proportions of readmitted in final data set*

***Age***

The next variable explored was age, which is a continuous variable representing 10-year intervals. Each class has a width of 10, indicating the *age* ranges covered by each interval. Given the structure of age, assessing its relationship with the logit of the outcome poses a challenge. To address this,age was transformed into a numeric variable to enable a clearer quantitative assessment of the relationship to *readmission*.

This transformation involved assigning the midpoint value of each *age* group as its representative numeric value. For instance, if an individual fell into the *age* category of 30-40, their numeric value would be set to 35; likewise, for the category of 60-70, the numeric value would be assigned as 65. These newly assigned numeric values allowed for the evaluation of the linearity of *age* with respect to the logit of the outcome. Examining the histogram of age*,* it appears the distribution is somewhat normal, with a slight left skew.

**

*Figure 4: Distribution of* age *in final data set*

*Table 3: Variable summary of* age *in final data set*

Min. 1st Qu. Median Mean 3rd Qu. Max.

15.00 55.00 65.00 65.34 75.00 95.00

By utilizing the midpoint of each class to represent *age* categories, we observe that the minimum age recorded is 15 years old. However, it's crucial to note that these values correspond to the midpoint of the 10-20 *age* category. In the finalized dataset, all encounters related to pediatric cases were excluded, along with ages ranging from 0 to 10 years. Nevertheless, ages falling within the 10-20 range were retained because they provide pertinent information about the adult population.

The first quartile, representing the 25th percentile, indicates that a quarter of the individuals are aged 55 or younger. The median age, which marks the midpoint of the dataset, stands at 65 years, suggesting that half of the individuals are 65 or younger. The mean age, calculated at 65.34 years, provides a measure of central tendency for the dataset's *age* distribution. Furthermore, the third quartile, at 75 years, implies that three-quarters of the individuals are aged 75 or younger. Finally, the maximum age recorded is 95 years old, indicating the presence of older individuals within the dataset. Overall, we observe that this population is relatively diverse in terms of age, however there is a notable concentration of individuals between the ages of 55 and 75. See *Table 3* for a variable summary of age*.*

***Gender***

The next variable to be explored was *gender*. There were 3 missing values for this variable in the original data set. Levels of either *male* or *female* were assigned to these records randomly with no evidence that systematic changes were introduced among the distributions of *gender* or *readmitted*. As these three records make up less than 0.005% of the entire data set, it was assumed that any potential impact on the distributions of *gender* or *readmitted* due to randomly assigning levels to the missing values is negligible. To support this assertion, a 2-sample test for equality of proportions with continuity correction was conducted to determine whether there was a significant difference in the distribution of *gender* before and after assigning random levels to missing values.

To conduct this test, counts were defined for males and females in each group, representing the number of individuals belonging to each level of *gender*. Then, the *prop.test()* function in R is used to perform the test. The test of two proportions yielded a test statistic of 1.9666e-27 with 1 degree of freedom, resulting in a *p-value* of 1. With such a high *p-*value, there is insufficient evidence to reject the null hypothesis of no difference in proportions of males between the two groups. Therefore, we fail to find statistically significant evidence that the proportions of males differ between the two groups. It can be concluded that randomly assigning a level of *gender* to the three missing values of *gender* did not introduce any systematic biases in the data. The distribution of *gender* in the final data set is visualized through a bar chart in *Figure 5*.

2-sample test for equality of proportions with continuity correction

data: c(group1\_male, group2\_male) out of

c(group1\_male + group1\_female, group2\_male + group2\_female)

X-squared = 1.9666e-27, df = 1, *p-*value = 1

alternative hypothesis: two.sided

95 percent confidence interval:

-0.004335559 0.004347623

sample estimates:

prop 1 prop 2

0.4623690 0.4623629

*Table 4: Results of 2-sample test for equality of proportions with continuity correction following imputation of the gender variable*

A graph with blue and green squares

Description automatically generated

*Figure 5: Distribution of gender in final data set*

***Medical Specialty***

The variable *specialty* represents the medical specialty of the physician or practitioner by which the patient was referred for their encounter. In the original dataset, *specialty* encompassed 73 unique values, with the predominant category being missing values, accounting for 49,949 records. Referencing the documentation provided by Strack et al. (2014), a comprehensive spreadsheet detailing associated codes and corresponding medical specialty names facilitated the translation of codes and categorization of variable levels. After careful consideration, it was determined that dividing medical specialty into the categories outlined in *Figure 6,* achieved a balanced representation of the data while ensuring interpretability. Despite exploring various categorical levels, medical specialty did not demonstrate significant predictive power in subsequent logistic regression models. Consequently, *specialty* was excluded from the model, likely attributable to the extensive presence of missing data.

A graph showing different colored squares

Description automatically generated

*Figure 6: Bar chart of medical specialty*

***Diagnosis Codes***

ICD-9 codes are a standardized method for categorizing and encoding diagnoses, symptoms, and medical procedures within the healthcare field. These codes, consisting of alphanumeric characters, are organized into chapters and sections based on body systems and medical conditions, ranging from values 0 – 999 with over 14,000 unique available codes. Each code provides a shorthand representation of a specific disease, injury, symptom, or medical procedure encountered in clinical practice. For each patient encounter in the data, three diagnoses codes were provided (*diag\_1, diag\_2, diag\_3*). These represent the primary, secondary and tertiary diagnoses at the encounter, in the form of ICD-9 codes. There were 665 unique codes in *diag\_1*, 614 unique codes in *diag\_2* and 632 unique codes in *diag\_3*. In the original data set, these diagnosis codes are either three- or five-digit ICD-9 codes. Data preprocessing was used to categorize the ICD-9 codes. See *Table 5* for summary of *diag\_1.*

The utilization of ICD-9 codes served two primary purposes in our analysis: first, for generating comorbidity scores, and second, for establishing a method to monitor diagnoses among patients with high comorbidity scores. By preserving the original diagnosis variables within our dataset, we can effectively evaluate the subset of patients with elevated comorbidity scores and discern the specific diagnoses associated with this group. This approach allowed us to investigate whether these patients exhibited similar or differing diagnoses compared to commonly recognized related diseases, such as renal disease, heart problems, retinopathy, and kidney disease.

***Discharge Disposition***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group Name | ICD-9 codes | Number of encounters | % of encounters | Description |
| Circulatory | 390-459, 785 | 8,849 | 23.27% | Diseases of the circulatory system |
| Respiratory | 460-519, 786 | 3,476 | 12.68% | Diseases of the respiratory system |
| Diabetes | 250.xx  250.x3 and 250.x2  250.x1 | 4,698  3,064  995 | 12.36%  8.06%  2.62 % | Diabetes (Unspecified)  Uncontrolled Diabetes Mellitus  Controlled Diabetes Mellitus |
| Digestive | 520–579, 787 | 3,043 | 8.00% | Diseases of digestive system |
| Musculoskeletal | 710-739 | 2,307 | 6.07% | Diseases of the musculoskeletal system |
| Genitourinary | 580-629, 788 | 1,419 | 3.73% | Diseases of genitourinary system |
| Other  (20.7%) | 800-999  680-709, 782  780,781,784,790,799  290-319  140-239  240-279  1:139  630-679  E-V  320-359  280-289  360-389  740-759 | 1,746  1,203  977  966 929  901  565  446  345  313 295 120  20 | 4.59%  3.16%  2.57%  2.54%  2.44%  2.36%  1.49%  1.17%  0.91%  0.82%  0.78%  0.32%  0.05% | Injury and poisoning  Diseases of the skin and subcutaneous tissues  Other symptoms and ill-defined conditions  Psychological disorders  Neoplasms  Other endocrine and metabolic diseases  Infections and parasitic diseases  Childbirth/pregnancy complications  External factors of injury  Diseases of the nervous System  Diseases of the blood and blood-forming organs  Diseases of the sense organs  Congenital anomalies |

*Table 5: Summary of primary diagnosis (diag\_1)*

Discharge disposition refers to the subsequent plan for a patient’s ongoing care or management prior to their hospital stay. In the original dataset*, discharge disposition* is represented by numeric codes, comprising 16 unique values. See *Table 5* for a summary of *discharge disposition* in the final data set. The accompanying documentation provided by the authors facilitated the translation of these codes into meaningful categories. The distribution of discharge disposition is illustrated in *Figure 7*.

Patients who were discharged to home make up 59.139% of records, with the second largest category being a skilled nursing facility (SNF) at 13.734%. To address the challenge of class imbalances and simplify subsequent analyses, a binary variable named *discharge* is derived. This binary variable categorizes patients into two groups: those discharged to home and those discharged elsewhere (including SNFs and all other facilities).

By collapsing multiple discharge disposition categories into a binary outcome, this variable facilitates more straightforward modeling and interpretation while still capturing the most impactful conclusions about discharge outcomes and transitions in patient care (see *Figure 8)*. Although it could be argued that some complexity is lost, updating this variable is guided by goals of parsimony and supports the goals of building a model with sufficient predictive power. Converting *discharge disposition ID* to a binary variable decreases model accuracy from 66.52% to 62.38% but increases model sensitivity from 55.20% to 58.62%. Due to the increased sensitivity, *discharge* was retained a contending predictor variable in its binary form.

A graph with numbers and text

Description automatically generated with medium confidence

*Figure 7: Distribution of discharge disposition ID*

A green and pink rectangles

Description automatically generated

*Figure 8: Distribution of discharge disposition ID in final data set*

|  |  |
| --- | --- |
| *Table of counts: discharge* | |
| Home | Other |
| 29918 | 17449 |

*Table 6: Distribution of discharge (binary) in the final data set*

***Admission source ID***

*Admission source ID*, indicating where a patient was admitted from, underwent a similar data preprocessing approach as *discharge disposition ID*. In the original dataset, there were 17 unique categories for admission source, represented by numeric codes. See *Figure 9* for a distribution of admission source ID in the final data set. The accompanying documentation provided by the authors facilitated the translation of these codes. Interestingly, 15 of the 17 categories represented less than 1% of the dataset each. Notably, the category *ER* accounted for a significant portion, comprising 56.496% of the dataset, followed by physicians' referral at 29.052%. The third largest category was *missing* while transfer from hospital represented only 3% of the data.

To streamline the variable and improve interpretability, all transfers—including those from hospitals, skilled nursing facilities, other facilities, critical access hospitals, and home health agencies—were grouped under the category *transfer*. Similarly, all referrals, originating from physicians, clinics, HMOs, or court orders, were consolidated. The category *missing* encompassed instances of missing data and other unspecified entries. The category *ER* was inclusive of all emergency room visits, as well as transfers from ambulatory centers and critical access facilities. Overall, assessing these distributions revealed that most patients were admitted based on a referral from a physician, clinic or health maintenance organization (HMO).

A graph of a number of colored squares

Description automatically generated

*Figure 9: Distribution of admission source ID in final data set*

***Admission type ID***

***A graph of a number of colored squares

Description automatically generated***

*Figure 10: Distribution of Admission Type ID in final data set*

*Admission type ID*, denoting the method through which a patient was admitted, underwent a similar preprocessing procedure to *admission source ID*. In the original dataset, there were 8 unique values for admission type, represented by numeric codes. The accompanying documentation provided by the authors facilitated the translation of these codes. Notably, 5 of these values or categories constituted less than 1% of the original data each. Among the predominant categories, *emergency/trauma* emerged as the largest, accounting for 53.05% of the original dataset, followed by urgent at 18.159% and elective at 18.542%. Levels such as NULL, not mapped, and not available were grouped together as *missing*. This consolidation allowed for a more cohesive representation of missing or unspecified entries within the dataset. Overall, the analysis highlighted the prevalence of emergency and urgent admissions, indicating the significance of these categories in the admission process***.*** See *Figure 10* for bar chart of admission type.

***Medication Type***

There are 23 columns in the dataset that indicate whether a patient is taking specific diabetic medications. Most of these columns exhibit significant imbalance, with most of these variables having less than 5% of all patients prescribed the given medication. See the glossary on page 121 for a list of these medications. To address the imbalances and consolidate important variable information, a new variable, *medication type,* was created to categorize patients based on their medication regimen. The objective is to streamline the dataset by differentiating between patients prescribed oral medications, those receiving injectable insulin, and those on a combination of both medication types. This consolidation aims to simplify the data while retaining the most important information about patients' medication regimens. See *Figure 11* for the distribution of *medication type.*

Lists of the oral and injectable medications were defined, using the column fields in the data set as oral medication and insulin as an injectable. Conditional statements were employed to determine if patients are on oral medications or injectable insulin, with their medication status being categorized accordingly. If a patient is found to be taking both oral and injectable medications, they are classified as being on both types of medication.

***A graph showing different colored rectangles

Description automatically generated***

*Figure 11: Distribution of Medication Type in final data set*

***Time in Hospital***

The total number of days a given encounter lasts is represented by *time in hospital*. A histogram was constructed to assess the distribution of this variable. See *Figure 12.* The right-skewed distribution indicates that most encounters tend to have relatively brief hospital durations, with a few instances of significantly prolonged stays, observed at the upper end of the spectrum.

The minimum recorded time spent in the hospital is 1 day, indicating that some individuals had very brief hospitalizations. Moving through the quartiles, we find that 25% of patients spent 2 days or fewer, while the median duration of hospitalization stands at 3 days. This implies that half of the patients had stays lasting 3 days or fewer. The mean time in hospital is calculated to be approximately 4.24 days, suggesting a slightly longer average duration compared to the median.

Progressing to the third quartile, we observe that 75% of patients had hospital stays of 6 days or fewer. The maximum recorded duration is 14 days, indicating that some patients experienced relatively extended hospitalizations. Interestingly, while the average time suggests a moderate overall duration, the presence of patients with stays as short as 1 day and as long as 14 days underscores the variability in hospitalization lengths within the dataset. *Table 7* shows the variable summary of *time in hospital.*

***A graph of a number of patients

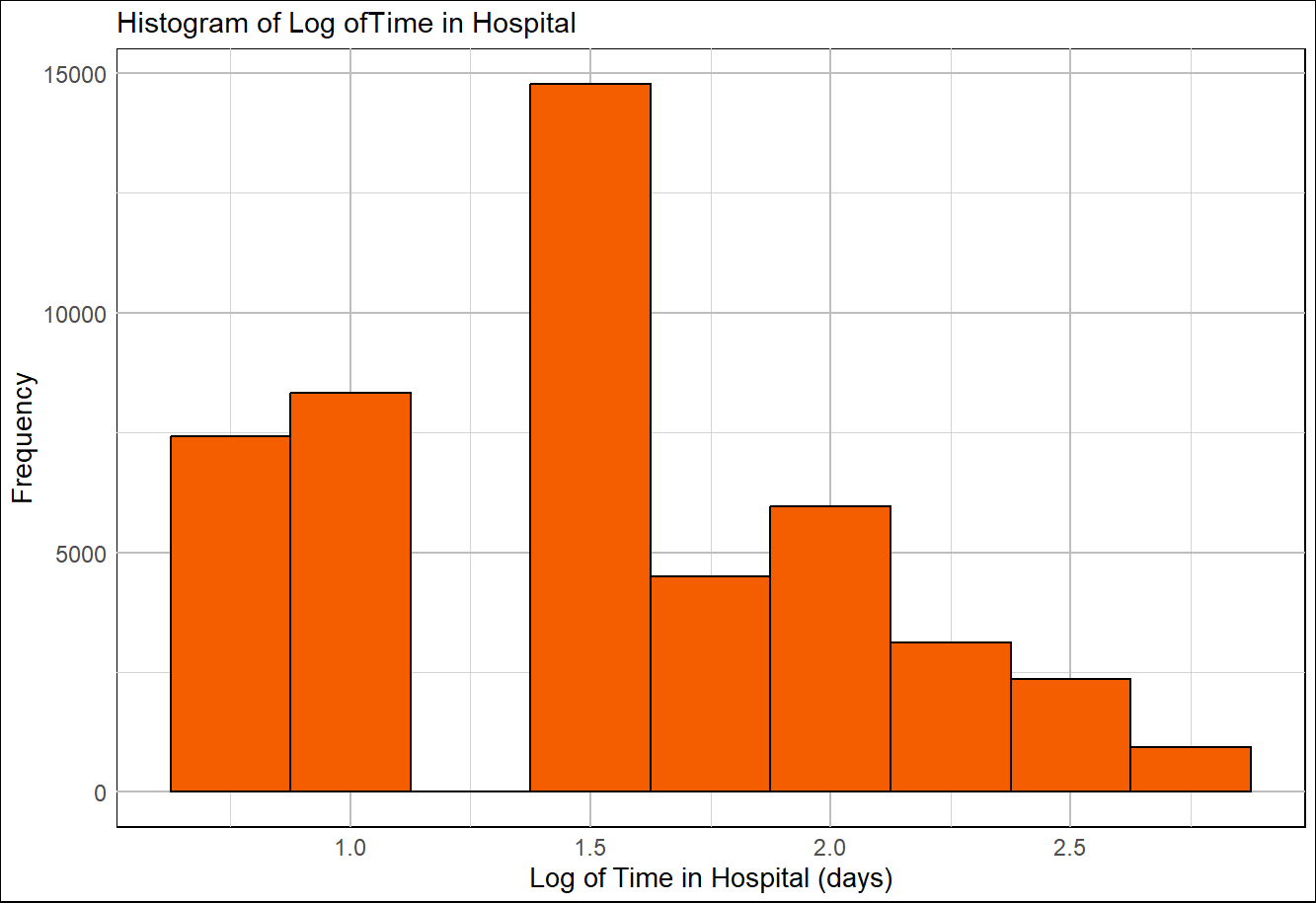
Description automatically generated***

*Figure 12: Distribution of Time in Hospital in final data set*

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.000 2.000 3.000 4.238 6.000 14.000

*Table 7: Variable summary of Time in Hospital in final data set*



*Figure 13: Distribution of Log of Time in Hospital in final data set*

To address the right-tailed nature of *time in hospital* and assess whether transformations could make it more suitable for modeling purposes, various techniques were applied. These included standardization, logarithmic, Box-Cox, and square root transformations. A log transformation resulted in the most optimal performance in the logistic regression model. Compared to the model using the untransformed *time in hospital,* accuracy reduced slightly from 62.96% to 62.43% but sensitivity increased from 58.211% to 58.46%. While only a slight improvement in model sensitivity, the log transformation produced the optimal results of *time in hospital* for modeling goals. See *Figure 13* for a histogram of the log-transformed ­*time in hospital* variable in the final data set.

***Number of Medications***

*Number of medications* represents the total number of medications a patient is prescribed. See *Figure 14* for a histogram of *number of medications.* The right-skewed distribution indicates that most patients are prescribed 10-20 medications, with fewer instances of larger quantities observed at the upper end of the spectrum. Potential high-leverage points appear at the end of the right tail. These points could potentially exert considerable influence on model outcomes due to their extended durations, warranting careful consideration during data analysis and interpretation. The largest number of medications prescribed to a patient is 81, but a median of 15 suggests that a value that high is uncommon.

***A green and black line graph

Description automatically generated***

*Figure 14: Distribution of Number of Medications in final data set*

The variable summary for *number of medications* (*see Table 8*) provides further insights into medication regimen complexity of patients within the dataset. The minimum recorded number of medications is 1, indicating that some patients were prescribed a single medication. As we progress through the quartiles, we observe that 25% of patients were prescribed 10 or fewer medications (first quartile), while the median number of medications stands at 14. This implies that half of the patients were prescribed 14 medications or fewer, serving as a central reference point for medication volume. The mean number of medications is calculated to be approximately 15.62, suggesting a slightly higher average than the median, indicating a tendency towards more extensive medication prescriptions within the dataset.

Advancing to the third quartile, we find that 75% of patients were prescribed 20 medications or fewer, highlighting the prevalent trend of patients receiving relatively moderate medication volumes. However, the maximum recorded number of medications is notably high at 81, indicating instances where patients were prescribed a substantial number of medications, reflecting complex medical conditions or treatment requirements. This wide range, from 1 to 81 medications, underscores the considerable variability in medication prescriptions among patients.

To address the right-tailed nature of *number of medications* and assess whether transformations could make it more suitable for modeling purposes, various transformation techniques were applied. These included standardization, logarithmic, Box-Cox, and square root transformations. A log transformation resulted in the most optimal performance in the logistic regression model. Compared to the model using the untransformed *number of medications,* accuracy reduced slightly from 62.43% to 62.38% but sensitivity increased from 58.46% to 58.36%. While only a slight improvement in model sensitivity, the log transformation produced the optimal results of *number of medications* for modeling goals. See *Figure 15* for a histogram of the log-transformed ­*number of medications* variable in the final data set.

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 10.00 14.00 15.62 20.00 81.00

*Table 8: Variable summary of number of medications in final data set*

A graph of a number of medications

Description automatically generated

*Figure 15: Distribution of Log of Number of Lab Procedures in final data set*

***A graph of a number of lab procedures

Description automatically generatedNumber of Lab Procedures***

*Figure 16: Distribution of Number of Lab Procedures in final data set*

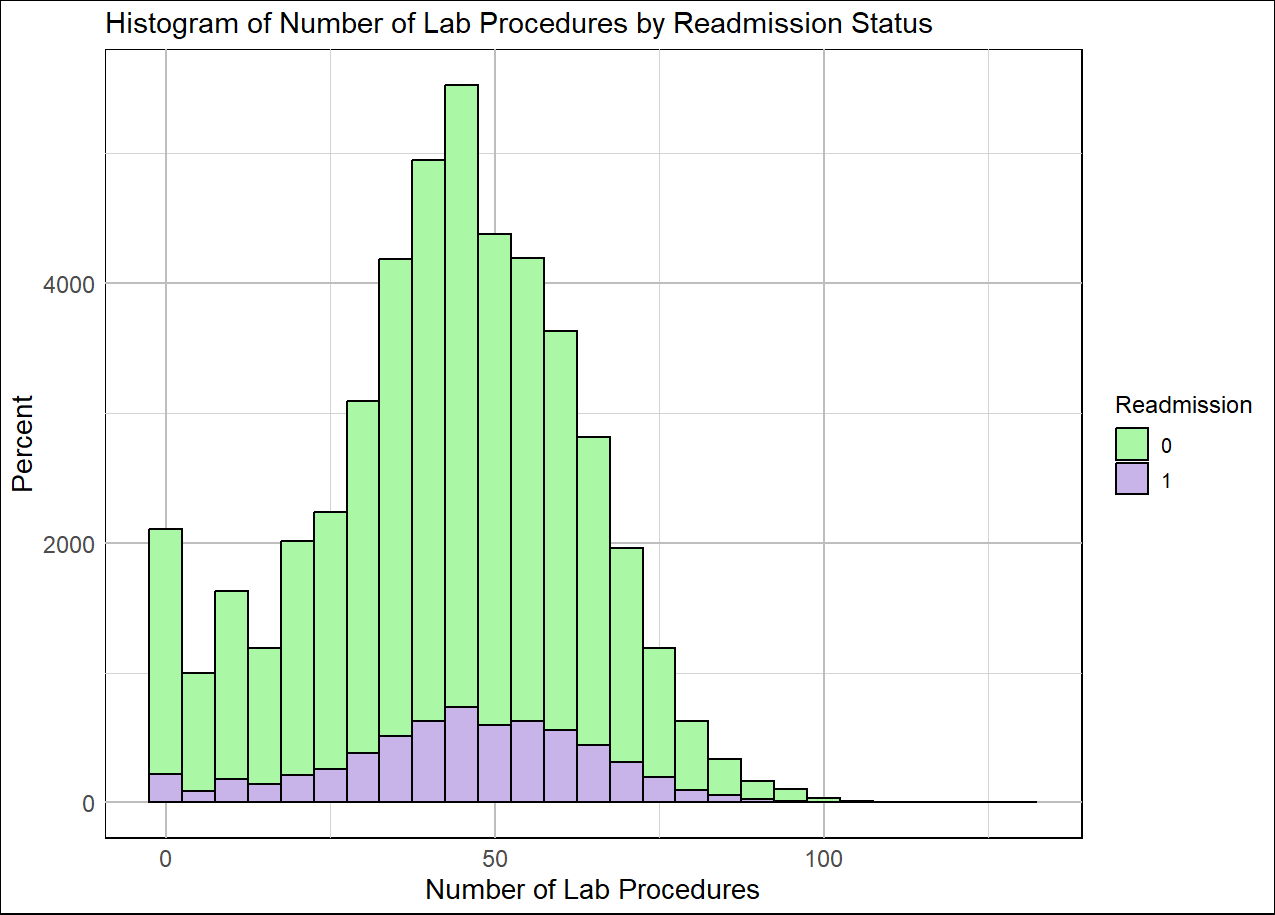
The variable *number of lab procedures* represents the total number of lab procedures a given patient has during their encounter. The distribution is relatively normal, despite the large spike at the *y* = 1 mark, indicating most patients had one lab procedure during their encounter. Potential high-leverage points appear at the end of the right tail. The largest number of lab procedures performed on a patient is 132, but with a mean value of 44, this is nearly triple the average number. This predictor appears to exhibit an uneven distribution characterized not only by its long right tail but also by a substantial spike in values preceding the left tail.

Min. 1st Qu. Median Mean 3rd Qu. Max.

1.0 31.0 44.0 42.8 57.0 132.0

*Table 9: Variable Summary of Number of Lab Procedures in final data set*

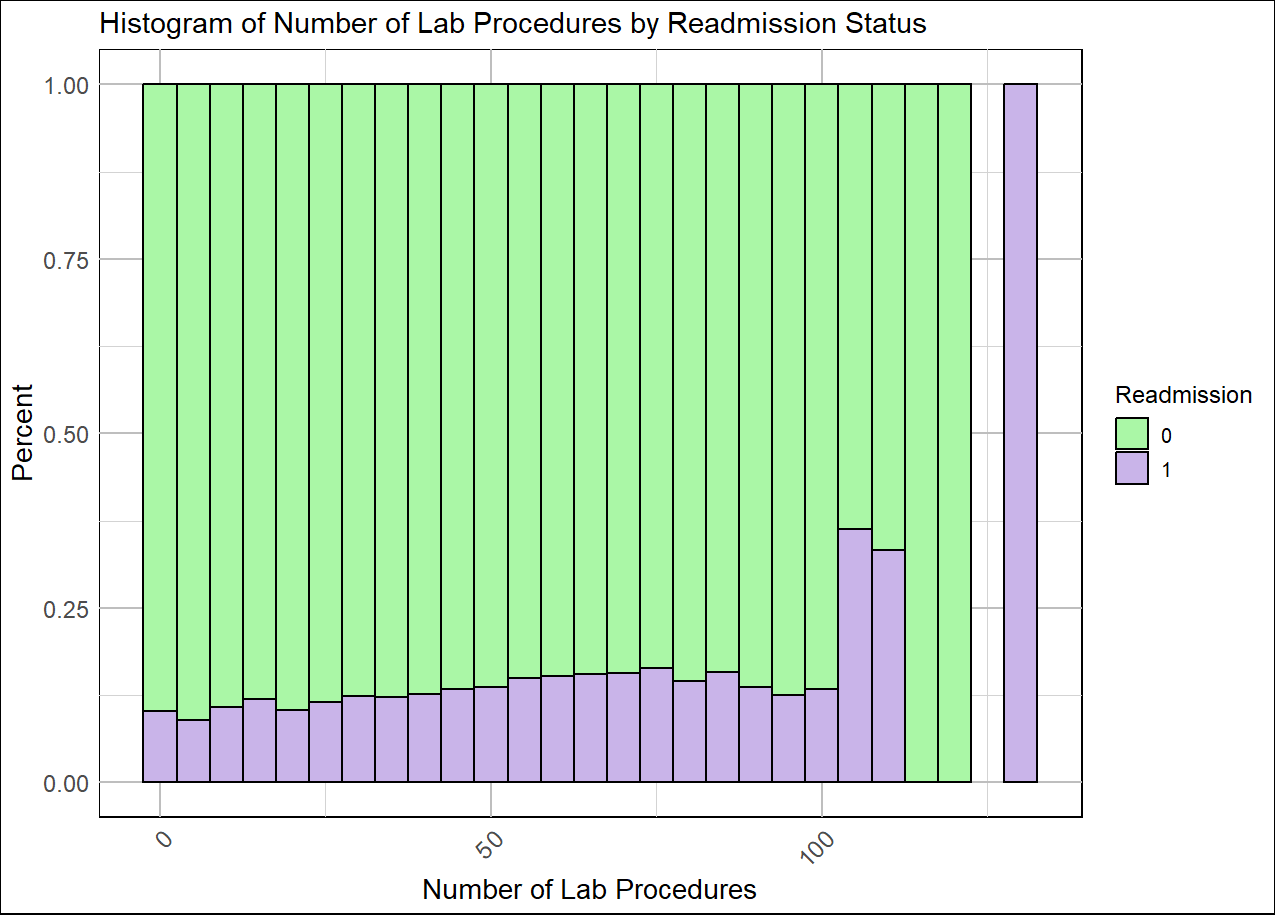
Upon reviewing both the normalized histogram (*Figure 17*) and non-normalized histogram (*Figure 18*) of the number of lab procedures with readmission overlay (binned by 5), we observe a slight positive trend: as the number of lab procedures increases, the



*Figure 17: Histogram of number of lab procedures with readmission overlay*

readmission rate also tends to increase. This trend becomes noticeable around the y = 20 mark and continues to rise until reaching y = 132. However, it's important to note that although there is a clear increase in readmission rates after 100 lab procedures, the number of observations beyond this point is considerably smaller. Consequently, selecting 100 as a cutoff value for the derived binary variable may not yield practical utility. As a result, cutoff values closer to the mean were explored.

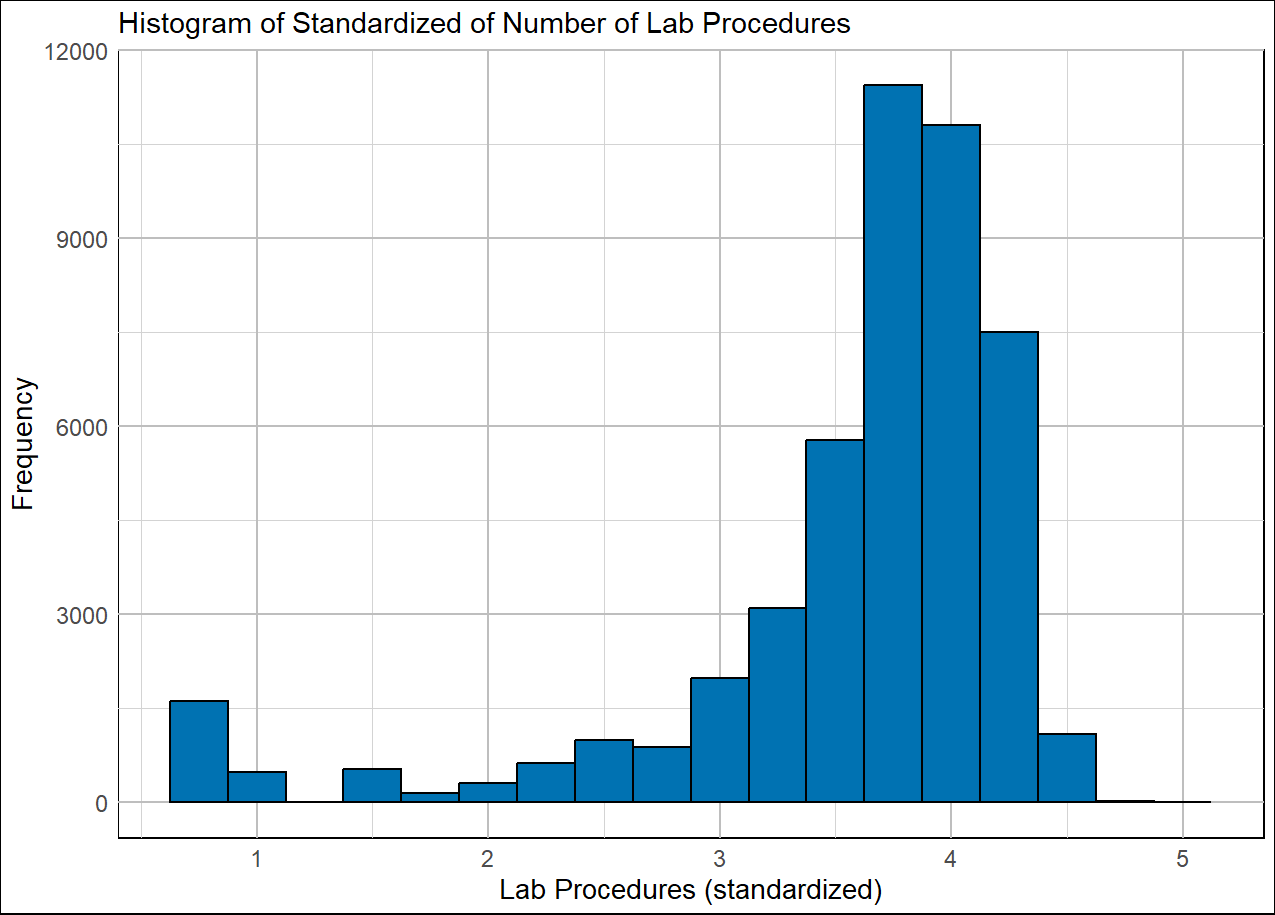
The number of lab procedures was transformed into a binary variable using various cutoff points, including patients with more than 1, 20, and 40 lab procedures. While this transformation helped address the issue of unbalanced distribution, it did not



*Figure 18: Normalized Histogram of number of lab procedures with readmission overlay*

lead to improvements in predictive performance. Interestingly, some of these transformations resulted in identical sensitivity outcomes.

Without variable transformation, the model accuracy using *number of lab procedures* was 62.42%, with a sensitivity of 58.537%. Converting the variable to a binary with a cutoff of greater than 1 lab procedure slightly improved model accuracy to 62.48%, but reduced sensitivity to 58.293%. Increasing the cutoff value to 20 further decreased model sensitivity to 57.886%. Surprisingly, raising the cutoff to 40 increased model sensitivity back to the baseline of 58.537%, indicating no discernible improvement over the untransformed variable.

**

*Figure 19: Histogram of number of lab procedures under log transformation*

Other variable transformations were explored to manage the non-normal distribution of *number of lab* procedures. These included standardization, square root transformation, Box-Cox transformation and logarithmic transformations. Standardizing *number of lab procedures* increased accuracy slightly to 62.52% and sensitivity to 58.537%, while the log transformation reduced accuracy to 62.32% and improved sensitivity to 58.293%. Box-Cox and square root transformations yielded the most optimal performance in the logistic regression model as far as sensitivity, increasing it to 58.618%. However, the Box-Cox transformation outshines the square root transformation as it improves accuracy too, to 62.57% while the square root transformation leads to a slight decrease in accuracy at 62.36%.

Despite the enhanced performance observed with the Box-Cox transformation, this improvement was likely influenced by its increased prominence in the dataset, attributed to the transformation expanding its range from 0 to 132 to 0 to 8000. To mitigate this disparity in scale, the variable underwent standardization following the Box-Cox transformation. However, this adjustment did not yield the desired outcome; instead, it led to a decline in both accuracy and sensitivity. Furthermore, although the Box-Cox transformation resulted in the optimal model performance, integrating it into the model caused the coefficients of other predictors to become exceedingly small. This result suggests that the transformed variable was being disproportionately weighted compared to others, thus contributing to the model's improved performance.

Although the distribution of *number of lab procedures* under the tested transformations does not achieve complete normality, ultimately the log-transformed variable was retained in subsequent modeling steps. Using this transformation allows for the largest number of positive occurrences to be identified, supporting the goals of the analysis and practical utility of the model by identifying the largest number of positive records. Refer to *Figure 19* for a histogram of the Box-Cox transformed version of *number of lab procedures.*

***Hemoglobin A1c Result***

*A1Cresult* is categorical variable that indicates whether a patient’s hemoglobin bA1c levels were tested and what the levels were. *A1Cresult* was explored for goals beyond maximizing predictive power, but for broader implications in understanding patient characteristics and the impact of this test on long-term outcomes. By exploring the association between HbA1c and other predictor variables, hypotheses for future investigations can be formulated. See *Figure 20*  for a distribution of *A1Cresult* in the final data set.

*A1Cresult* levels include *Norm* indicating a normal test score, *>7* indicating levels higher than 7 but less than 8, *>8* indicating levels higher than 8, and *None* indicating no test was taken. Most patients did not receive HbA1c tests, for a total of 35,216 or 80.941% of the records. The next largest category is *Norm* at 8.915%, *>8* at 5.782% and *>7* at 4.360%. A table of A1Cresult counts and proportions was constructed to summarize the distribution of HbA1C results (see *Table 10* and *Table 11*).

*A graph with multiple colored bars

Description automatically generatedFigure 20: Bar chart of Hba1C results in final data set*

None >8 Norm >7

35,216 3,879 2,516 1,897

None >8 Norm >7

80.941% 08.915% 05.782% 04.360%

*Table 10: A1Cresult totals* *Table 11: A1Cresult proportions*

Upon examining the contingency table of readmitted against A1Cresult (*Table 12*), we can see distinct distributions of HbA1c test results across the readmission status. Among patients who were readmitted, the proportions of different HbA1c levels

*Table 12: A1Cresult against Readmitted in contingency table with column proportions*

exhibit slight variability, ranging from approximately 12.10% to 13.43%. Notably, the proportion of readmitted patients among patients who did not receive an Hba1C test is higher compared to those who did receive testing, regardless of the result of the test. Additionally, patients with normal test results had lower readmission rate (12.039%) than patients with results scoring >7 (12.36%) and >8 (12.30). Interestingly, patients with A1c results higher than 7 exhibit higher readmission rates than patients with scores higher than 8.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *A1Cresult* against *readmitted* with column proportions | | | | |
|  | >7 | >8 | None | Norm |
| 0 | 87.644% | 87.705% | 86.565% | 87.961% |
| 1 | 12.356% | 12.295% | 13.435% | 12.039% |

Pearson's chi-squared test for independence was used to determine whether there is a significant association between HbA1c *result* and *readmission* (*s*ee *Table 13*)*.*  The null hypothesis (H0) for this test is that there is no association between the two variables, meaning that the proportions of readmitted individuals are independent of the HbA1c levels. The alternative hypothesis (H1) is that there is an association between the variables, indicating that the proportions of readmitted individuals differ across different HbA1c levels. In this case, with a test statistic of X-squared = 9.16 and *p-*value = 0.0272, we have enough evidence to reject the null hypothesis and assume that there is a significant association between *A1Cresult* and *readmission*.

Pearson's Chi-squared test

X-squared = 9.1599, df = 3, p-value = 0.02724

*Table 13: Pearson’s Chi-squared test of independence for readmitted against A1Cresult*

***Diagnosis Codes***

Strack et al. (2014) observed that the association between the likelihood of readmission and HbA1c measurement is influenced by the primary diagnosis. Patients with a primary diagnosis of diabetes were more inclined to undergo HbA1c testing compared to those with primary diagnoses of circulatory or respiratory conditions. However, the authors did not provide specific details regarding the type of diabetes or any additional information about the diagnosis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Percentage of Readmissions against HbA1c result | | | | | |
| Primary Diagnosis | >7 | >8 | None | Norm | Total % Readmitted |
| Diabetes (unspecified) | 5.49% | 17.99% | 69.46% | 7.06% | 100% |
| Diabetes (uncontrolled) | 1.93% | 39.77% | 53.64% | 4.65% | 100% |
| Diabetes (controlled) | 3.55% | 25.44% | 66.27% | 4.73% | 100% |
| Respiratory | 5.71% | 8.61% | 78.91% | 6.77% | 100% |
| Circulatory | 4.81% | 9.01% | 80.17% | 6.01% | 100% |

*Table 14: Percent*age *of readmissions against HbA1c result per primary diagnosis*

In the dataset, the only ICD-9 codes longer than 3 digits were those for diabetes (250.XX), which offer additional characteristics about the diagnosis. The cleaned ICD-9 codes were utilized to subset the patients based on whether their diabetes type is controlled (codes ending in 250.X1), uncontrolled (codes ending in 250.X2 and 250.X3), or unspecified (all other 250.XX codes). Using the categorized diagnoses, HbA1c results were explored of patient subgroups with different primary diagnoses.

*Table 14* provides insights into the distribution of patients across various categories of HbA1c results based on their primary diagnosis. Patients with diabetes, regardless of whether it was specified as controlled, uncontrolled, or unspecified, had a higher proportion of patients receiving HbA1c tests compared to patients with a primary diagnosis of circulatory or respiratory. This parallels with the results of Strack et al. (2014) who observed this association.

Among the different subsets of diabetes diagnoses, we see patients with uncontrolled diabetes were tested the most frequently, with only 53.64% not receiving tests compared to 66.27% of patients with controlled diabetes and 69.56% of patients with unspecified diabetes. These findings underscore differences in HbA1c testing patterns among different diabetes subsets and suggest that healthcare providers may prioritize testing for patients with uncontrolled diabetes over other conditions.

***Comorbidity scores***

To explore the relationship of readmission to HbA1c result with respect to patient diagnoses further, HbA1c result and patient comorbidity scores were assessed through a series of statistical tests. Comorbidity scores offer a more comprehensive evaluation of patients' health status by incorporating all diagnoses, including secondary and tertiary conditions. See page 61 for how these scores are constructed. Additionally, these scores apply weights to each diagnosis based on their severity, providing a nuanced understanding of patients' overall health complexity. This approach enables the identification of high-risk patient subgroups that may be overlooked when considering primary diagnoses alone.

Assessing the boxplot (*Figure 19*) of *scores* against *readmission*, several key observations emerge. First, the distribution of means and medians across the different readmission categories appears to be relatively uniform and clustered around zero. This is largely influenced by a significant portion of patients having a score of 0, indicating a baseline or minimal health issue. However, upon closer inspection, it becomes evident that patients whowere not readmitted exhibit a wider range of scores, with a maximum score of about 9.5. In contrast, all other readmission categories do not have scores that range that high.

This finding is intriguing because conventional wisdom might imply that higher scores, reflecting the presence of higher risk conditions and poorer health, would correspond to a greater risk of early readmission. However, the data shows that among patients who were not assessed for their scores, a wider range exists, suggesting a greater presence of individuals with higher-risk health conditions. Therefore, the broader spectrum of scores observed among patients who did not undergo testing suggests the possibility of a less thorough assessment and management of their health conditions. This could potentially contribute to higher-risk health outcomes and an increased likelihood of early readmission.

The ANOVA results reveal a significant effect of HbA1c result on scores, indicating that there are notable differences in mean scores across the various levels of HbA1c result. The small p-value (1.92e-07) associated with the F-statistic (11.36) suggests that these differences are unlikely to have occurred by chance. With a degree of freedom of 3 for HbA1c result, the analysis highlights the presence of four distinct categories within this variable. The observed variation in scores attributed to differences

***A graph with red and blue squares

Description automatically generated***

*Figure 21: Boxplot of Charlson scores by HbA1c result*

in HbA1c result (sum of squares = 33) underscores the importance of glycemic control in influencing the measured outcomes. While further investigation is needed to elucidate the precise nature of these differences and their clinical implications, these findings underscore the potential impact of HbA1c levels on the observed scores, warranting attention in patient management strategies.

The Kruskal-Wallis test, a non-parametric alternative, also corroborates the finding of significant differences among the HbA1c result categories (χ²(3) = 23.126, p < 0.001). These results collectively suggest that HbA1c result categories significantly influence the CCI scores, with notably lower scores associated with higher HbA1c levels.

ANOVA

Df Sum Sq Mean Sq F value Pr(>F)

A1Cresult 3 33 10.946 11.36 1.92e-07 \*\*\*

Residuals 47363 45641 0.964

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*Table 15: ANOVA for CCI scores against HbA1c*

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = scores ~ A1Cresult, data = dfC)

$A1Cresult

diff lwr upr p adj

>8-None -0.13644289 -0.187730396 -0.08515539 0.0000000

Norm-None -0.02826716 -0.090829131 0.03429481 0.6516714

>7-None -0.04504120 -0.116497443 0.02641505 0.3676075

Norm->8 0.10817573 0.030571412 0.18578006 0.0019415

>7->8 0.09140170 0.006463921 0.17633947 0.0291229

>7-Norm -0.01677404 -0.108958733 0.07541066 0.9661721

*Table 16: Tukey multiple comparisons of means*

Kruskal-Wallis rank sum test

data: scores by A1Cresult

Kruskal-Wallis chi-squared = 23.126,

df = 3, *p-*value = 3.801e-05

*Table 17: Kruskal-Wallis rank sum test*

**Multicollinearity**

Multicollinearity occurs when predictor variables in a regression model are highly correlated with each other. This can cause unstable estimates of model coefficients and inflated standard errors. To assess multicollinearity within the data, a correlation matrix was produced (see *Figure 22*). A correlation matrix highlights the strength and direction of variable relationships. Each entry in the matrix denotes the correlation coefficient between two variables, ranging from -1 to 1. This coefficient quantifies the linear relationship where a value of 1 signifies a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 denotes no correlation. Correlation matrices aid in identifying multicollinearity and significant variable relationships which can be used to assist in variable selection.

***Correlation Matrix***

***A graph of a number of numbers and graphs

Description automatically generated with medium confidence****Figure 22: Correlation matrix of numeric predictors*

***Variance Inflation Factor***

VIF, or Variance Inflation Factor, is a commonly used metric to assess multicollinearity. VIF scores were calculated at various stages of the analysis, providing insights into the degree of correlation between predictor variables among different subsets. The VIF measures how much the variance of an estimated regression coefficient is inflated due to multicollinearity. Specifically, for each predictor variable in the model, the VIF quantifies how much the variance of its estimated coefficient is increased by the presence of multicollinearity with other predictor variables. The formula for VIF for a predictor variable is:

Where is the value obtained from regressing on all other predictor variables in the model. A VIF value of 1 indicates no multicollinearity, while values above 5 or 10 are considered problematic and may require further investigation.

VIF results of the final logistic regression model are in *Table 18*. Each row in the table represents a predictor variable, with columns indicating the GVIF (Generalized Variance Inflation Factor), the degrees of freedom (Df), and the GVIF adjusted for the number of predictors and their degrees of freedom. Adjusting the number of predictors and their degrees of freedom helps provide a clearer indication of multicollinearity among variables and prevent overfitting. These values help assess multicollinearity, where lower adjusted GVIF values suggest less multicollinearity among predictor variables.

By examining VIF scores, instances of potential multicollinearity were identified. Subsequently, VIF scores were utilized to refine the set of predictor variables included in the final model. This iterative process allowed the selection of a subset of predictors that demonstrated low multicollinearity, ensuring the robustness and accuracy of the logistic regression model.

Most VIF scores of the final logistic regression model are notably low, not exceeding 1.51, indicating minimal multicollinearity among the predictor variables. However, admission type ID and admission source ID stand out with VIF scores of 4.678 and 4.607, respectively. These values are approaching the threshold of 5, signaling potential multicollinearity issues. When either admission source ID or admission type ID was omitted from the model, VIF scores decreased to no more than 1.3 for either variable. However, this adjustment led to a significant decline in model performance with a 2.26% reduction in model sensitivity. Despite these variables approaching multicollinearity, they were retained in the model to serve analysis goals of identifying the largest number of positive records as possible.

GVIF Df GVIF^(1/(2\*Df))

race 1.053289 3 1.008690

gender 1.010730 1 1.005351

admission\_type\_id 4.674563 3 1.293074

admission\_source\_id 4.647064 4 1.211706

boxcox\_num\_lab\_procedures 1.351492 1 1.162537

log\_num\_med 1.516537 1 1.231478

preceding\_visits\_binary 1.020893 1 1.010392

change 1.393474 1 1.180455

diabetesMed 1.361436 1 1.166806

A1Cresult 1.125168 3 1.019850

log\_time\_hosp 1.411192 1 1.187936

log\_scores 1.012781 1 1.006370

F

*Table 18: Variance Inflation Factor scores for LR model*

**Data transformations**

Various data transformations were explored to enhance the utility of predictor variables. Data transformations come in various forms, each tailored to address specific challenges associated with the structure of the data. Some common transformation types include logarithmic transformations, normalization, polynomial transformation, and box-cox transformations. These transformations can prove useful by mitigating the effects of nonlinearity within the data, thereby improving the model's robustness and interpretability. Additionally, they aid in achieving better adherence to model assumptions, facilitating more accurate predictions.

Logarithmic transformations were employed to mitigate issues related to skewed distributions, facilitate a more uniform spread of data values, and achieve a normal distribution among continuous variables. The general form of a logarithmic transformation is: . Here, *x* represents the original data point, β is the logarithmic base and *y* is the transformed value. Logarithms transform multiplicative processes into additive ones where increasing *x* by a factor of β corresponds to adding 1 to *y*. This property is useful for visualizing data on a logarithmic scale, where each increment represents a multiplicative change rather than an additive one. Logarithmic transformation allows for a more intuitive interpretation of relationships between variables, as changes in the logarithmic scale correspond to proportional changes in the original scale.

Concurrently, standardization was utilized to ensure that all variables shared a consistent scale, allowing for a fairer comparative analysis. Also known as z-score normalization, standardization is a technique that transforms the scale of variables to have a mean of zero and a standard deviation of one. This process is particularly beneficial when dealing with features that are measured in different units or scales, as it puts all variables on a common scale, allowing for fair comparison and interpretation. By standardizing the data, outliers are also less influential, making statistical analyses more robust.

The process of standardization involves subtracting the mean of the variable from each data point and then dividing it by the standard deviation. The formula is as follows: , where *z* is the standardized value, *x* is the original data point, μ is the mean of the variable, σ is the standard deviation of the variable. This standardization of coefficients makes it easier to compare the relative importance of different predictors in the model.

The efficacy of each variable transformation was systematically evaluated through sensitivity analyses, where the impact of each transformation on model performance metrics was assessed. Ultimately, it was observed that logarithmic transformations yielded optimal results for variables *time in hospital and Charlson scores,* showing an increase in LR model sensitivity compared to the non-transformed version of these variables. The log-transformed versions of *time in hospital, number of lab procedures and Charlson scores* were retained as predictors in the model.

**Feature engineering**

Through the process of feature engineering, new variables were derived from existing data, utilizing domain knowledge and thorough understanding of the dataset. Through iterative experimentation, the most effective version of these derived variables were identified. Two additional features were created as predictors for the final model: *preceding visits* and *comorbidity scores*.

***Preceding visits***

The variables *number of outpatient visits, number of inpatient visits, and number of emergency room visits* within the past year represent related aspects of patient healthcare patterns. These variables are also measured on the same scale, where the values represent the total number of encounter types over the past year. These variables exhibit significant imbalances, with most values being 0 for each category, posing challenges for traditional assessments of linearity to the logit of the outcome.

Several techniques were experimented with to improve the usability of variables related to the number of inpatient, outpatient, and emergency room visits. Initially, a logistic regression model was built with all predictor variables except for *num emergency, num outpatient* and *num inpatient*. Then, a logistic regression model containing *num emergency, num outpatient* and *num inpatient* as independent predictors was built to show the impact of these variables. Adding these variables as independent predictors to the model only increases sensitivity by 0.1%.

Aggregating *num emergency, num outpatient* and *num inpatient* into a single metric, representing whether a patient had an inpatient, outpatient or ER encounter within the previous year was proven to be the most impactful approach for improving the predictive power of these variables. Different variations of *preceding visits* were experimented with, such as incorporating a weighting scheme, and converting it to a binary variable. See *Table 19* for model results*.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Metrics: LR Model using *preceding visits* | | | | | |
|  | Accuracy | Sensitivity | Specificity | AUC | F2 score |
| LR model base | 0.5937 | 0.59808 | 0.59325 | 0.595665739 | 0.343068 |
| num\_outpatient, num\_inpatinet, num\_emergency | 0.6123 | 0.58936 | 0.61452 | 0.60194 | 0.346986 |
| Preceding visits (total) | 0.6007 | 0.59024 | 0.60169 | 0.5959615 | 0.342230 |
| Preceding visits (binary) | 0.5973 | 0.59808 | 0.59721 | 0.5976461 | 0.34465 |
| Preceding visits (weighted) | 0.6066 | 0.59111 | 0.60815 | 0.59962 | 0.34531 |

*Table 19: Results of LR models with different trials using preceding visits variable*

The LR base model without using *num inpatient, num outpatient* and *num emergency* yielded an accuracy of 59.37%, sensitivity of 59.808%, specificity of 59.325%, and an AUC of 59.5665. When incorporating *num inpatient, num outpatient* and *num emergency*, accuracy improved slightly to 61.23%, but sensitivity decreased to 58.936%. However, when using *preceding visits binary,* a similar accuracy (59.73%) was achieved, and sensitivity increased to 59.808%. This higher sensitivity suggests that the binary representation effectively captures true positives, making it a favorable choice for identifying early hospital readmissions. Other representations of *preceding visits*, such as using the total number of preceding visits or constructing a weighting scheme, yielded slightly lower sensitivity values. Therefore, the LR model with *preceding visits* represented as a binary variable appears to offer the best balance of accuracy and sensitivity for predicting early hospital readmissions.

***Comorbidity scores***

An additional group of predictors that posed challenges for both model performance and interpretation are *diag\_1, diag\_2,* and *diag\_3*, which correspond to the ICD-9 codes indicating diagnoses recorded at each patient encounter. One notable advantage of the structure of these variables is that variable levels are the source ICD-9 codes, rather than pre-classified ones, allowing flexibility in categorizing the variables. Various strategies were employed to transform these variables into usable forms, including experimenting with different classifications of codes among the variable levels. The presentation of *diag\_1, diag\_2* and *diag\_3* by previous authors offers just one perspective on organizing this information, prompting exploration into alternative categorizations.

One method of organizing these codes involves transforming diagnoses into comorbidity scores, thereby expanding the analysis to encompass a wider range of factors. This approach not only enhances the predictive strength of these codes but also assigns meaningful rankings to patient diagnoses. Comorbidities, referring to the presence of additional medical conditions alongside the primary condition of interest, can affect patient outcomes. By utilizing these scores in the final model, we gain insights not just into patients' diagnoses but also into the magnitude of their impact or severity.

To quantify the impact of comorbidities, two widely utilized comorbidity scoring systems, namely the Charlson Comorbidity Index (CCI) and the Elixhauser Comorbidity Index (ECI), were used in this analysis. While the Charlson Comorbidity Index primarily considers conditions such as myocardial infarction, congestive heart failure, and diabetes, (for a total of 19 different diagnoses) the Elixhauser Comorbidity Index encompasses a broader spectrum of conditions spanning various organ systems (for a total of 38 diagnoses). These scoring systems assign numerical values to various comorbid conditions based on their severity and prognostic significance. Once the weights are assigned, the total Charlson score is calculated by summing the weights of all comorbid conditions present in a patient. The calculated scores provide a quantitative representation of the overall burden of comorbidities experienced by each patient.

Initially developed as a means to predict one-year mortality in medical patients based on the presence of 19 comorbid conditions, the CCI has seen significant changes over time. One notable aspect of its evolution is the expansion of the list of comorbid conditions included in the index. Subsequent versions have incorporated additional conditions that have emerged as significant predictors of mortality or other clinical outcomes, reflecting advances in medical knowledge and shifts in disease prevalence.

There have also been modifications to the weighting scheme used in the index, updated to better capture the relative impact of different comorbidities on patient outcomes. Additionally, researchers have adapted the CCI for use beyond in-hospital mortality, such as for various populations and clinical settings, like cancer patients, surgical patients, and specific disease cohorts. The broadening utility of comorbidity scores underscores their relevance in the model. The diagnoses and associated weights of the CCI are provided in *Figure 23.*

A table with a list of medical information

Description automatically generated

*Figure 23: Charlson Comorbidity Index Scoring*

To calculate the comorbidity scores for each patient, the *comorbidity()* function from the comorbidity package in R was employed. Initially, the dataset was restructured into a long format, maintaining encounter\_id as the identifier, and converting the diagnosis codes (diag\_1, diag\_2, diag\_3) into a single column. After reformatting, the comorbidity function was applied using the comorbidity package, with either the Charlson comorbidity or Elixhauser index specified. The resulting comorbidity scores were then calculated and appended to the original dataset. See *Table 20* for a variable summary of Charlson comorbidity scores.

To comprehensively evaluate the influence of CCI scores on both the dataset and model outcomes, normalized and non-normalized histograms were constructed. These histograms aimed to highlight any discernible relationships between readmission rates and CCI scores. In the normalized histogram, a subtle uptick in readmitted patients was observed, denoted by the upward trend in positive occurrences.

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 0.000 0.000 0.693 2.000 7.000

*Table 20: Variable summary of Charlson comorbidity scores*

*A graph of a bar chart

Description automatically generated*

*Figure 24: Normalized histogram of scores*

A graph of a bar graph

Description automatically generated

*Figure 25: Non-normalized histogram of scores*

A logistic regression model was trained using the different variations of the Charlson and Elixhauser scoring indexes using the *glm()* function. Models were evaluated using k-fold cross-validation and confusion matrices were constructed to assess model performance. Through a series of sensitivity analyses, it was determined that the most effective model includes the logarithm of Charlson scores. This optimization demonstrates how variables not only become more useful but are also consolidated into a single score with appropriate weights assigned to each diagnosis. This transition elevates these variables from mere factors (categories of diseases/diagnoses) to meaningful, systematically interpretable numeric outputs.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Metrics: LR Model with Comorbidity Scores | | | | | |
|  | Accuracy | Sensitivity | Specificity | AUC | F2 score |
| Base model | 0.603 | 0.56670 | 0.60660 | 0.5866 | 0.33139 |
| Elixhauser scores | 0.5983 | 0.58326 | 0.59979 | 0.59152 | 0.33737 |
| ln(elixhauser scores) | 0.6007 | 0.58152 | 0.59954 | 0.59052 | 0.33666 |
| Charlson scores | 0.5971 | 0.59198 | 0.59764 | 0.59480 | 0.34154 |
| ln(charlson scores) | 0.5937 | 0.59808 | 0.59325 | 0.59567 | 0.34306 |

*Table 21: Logistic regression model metrics using different comorbidity scores*

The LR base model yielded an accuracy of 60.3%, sensitivity of 56.67%, specificity of 60.66 %, and an AUC of 58.66. When incorporating Elixhauser and Charlson scores, model accuracy decreases slightly but sensitivity increases from 56.67% in the base model to 59.198% with Charlson scores and 58.326% with Elixhauser scores. However, when considering log transformations, we see the largest increase in sensitivity to 59.808% with the log of Charlson scores. This higher sensitivity suggests that the log transformed representation better captures true positives, making it the most favorable choice among these possibilities. The log of CCI scores was retained as a contending predictor in the model.

**Variable selection**

After establishing a set of candidate predictors, the following steps involve identifying the most influential variables. Implementing variable selection methods helps identify the optimal subset of predictors and refine contending predictive models. Leveraging techniques such assessing coefficient magnitudes, variable importance plots, correlation analysis, LASSO coefficients and stepwise variable selection procedures, predictors that showed the greatest influence on the readmission were identified. By prioritizing variables that offered the greatest predictive power while avoiding redundancy and multicollinearity, the most robust and parsimonious model can be made.

In health research models, the importance of variables extends beyond mere statistical significance and encompasses a range of factors such as clinical relevance, biological plausibility, reproducibility, and interpretability. Although variables such as demographic factors like *race* and *gender* may not consistently emerge as significant predictors in predictive modeling, their inclusion holds significant value as they enhance model interpretability and aid in adjusting for potential confounding effects.

Furthermore, in some cases, their incorporation is essential for legal and ethical compliance, ensuring fairness and transparency in the decision-making processes. In medical settings or research, decisions regarding the removal of certain variables from the model often require approval from various stakeholders, including ethics boards and research committees, to ensure adherence to ethical guidelines and regulatory standards. However, within the context of this analysis, the decision to retain a variable is guided by their relevance to the research question and their contribution to model performance.

***Stepwise Procedures***

Forward and backward stepwise variable selection are used to determine which variables are most important for predicting a given outcome. In forward stepwise selection, the procedure starts with an empty subset and iteratively adds one variable at a time. At each step, the variable that results in the best improvement in the model fit, measured using the AIC (Akaike Information Criterion), is added to the model. Conversely, backward elimination begins with a full model, where variables are removed one at a time, based on their contribution to the model fit as assessed by AIC. This process continues until no further improvement in model fit is observed, or until a predetermined stopping criterion is met.

With an AIC value of 79,450, the forward stepwise procedure concluded by retaining all variables from the initial subset. This suggests that none of the variables were considered less significant or redundant based on the AIC selection criteria. In the first step of the backward elimination process, *change* was removed, leading to a decrease in the AIC from 79,447.51 to 79,445.91. Subsequently, *gender* was removed in the next step, further reducing the AIC to 79,445.12. No additional variables were removed as their elimination did not lead to a substantial decrease in the AIC.

While the AIC decreased with the removal of variables in the initial steps, the decrease is not substantial enough to justify the exclusion of those variables from the model. Therefore, sensitivity tests were conducted to assess the impact of each variable individually on the model's sensitivity and specificity. This approach ensures that variables are not prematurely discarded based solely on their individual AIC contributions and allows for a more nuanced understanding of their importance in predicting the outcome of interest. Sensitivity tests indicated that while *gender* and *age* may not exhibit strong correlation with the outcome, there is not sufficient evidence to justify removing them from the model. Therefore, despite their limited predictive power, they were retained as contender predictor variables to assist in model interpretation.

***LASSO***

Similarly, regularization techniques like LASSO also aim to select the most impactful and important subset of predictor variables. Traditional regression methods can overfit the data or provide unstable estimates when faced with multicollinearity, where predictor variables are highly correlated. Regularization methods like LASSO (Least Absolute Shrinkage and Selection Operator) address these issues by introducing a penalty term to the regression equation, constraining the coefficients and shrinking them towards zero. This process encourages sparsity in the model, effectively selecting a subset of important predictor variables while disregarding less relevant ones.

The *glmnet* package was utilized for implementing LASSO regression. Using the balanced train data without high-leverage points and appropriate variable transformations, a train and test set of dummy variables were created. This dummified data set was used as input for the LASSO function. Logistic regression with LASSO regularization was executed using cross-validation to determine the ideal regularization parameter (lambda), regulating the extent of coefficient shrinkage. This approach aids in optimizing the model's performance by balancing the trade-off between model complexity and generalizability. Subsequently, the optimal lambda value is extracted, and the coefficients of the selected variables are obtained. This process enables the identification of important predictor variables while simultaneously penalizing less influential ones, facilitating model interpretation and potentially improving prediction accuracy.

Initially a model was fit with all predictor variables. Model coefficients of the variables obtained from the LASSO model were examined. These coefficients represent the strength and direction of the relationship between each predictor variable and the target variable. Variables with non-zero coefficients are considered important predictors, while those with zero coefficients have been effectively eliminated from the model. The variables associated with non-zero coefficients were retained in the model. LASSO coefficients of the logistic regression model containing all contending predictor variables are available in *Table 22*. One level of each of the categorical variables does not appear in these results, as one category is used as the reference level as part of the one-hot encoding process.

"Optimal lambda: 0.000658873983839763"

31 x 1 sparse Matrix of class "dgCMatrix"

s1

(Intercept) -1.633378822

race .

gender .

admission\_type\_id .

admission\_source\_id .

discharge .

log\_num\_lab 0.116691280

log\_num\_med 0.031601399

preceding\_visits\_binary 0.794909274

log\_scores -0.033507257

change .

diabetesMed .

A1Cresult .

log\_time\_hosp 0.229161248

race\_Caucasian 0.028726589

race\_Missing -0.230721634

race\_Other -0.191043798

gender\_Male 0.017698026

admission\_type\_id\_Elective -0.108336858

admission\_type\_id\_Emergency\_or\_trauma -0.117674594

admission\_type\_id\_Urgent 0.038472966

admission\_source\_id\_ER 0.145719215

admission\_source\_id\_Other 0.164337891

admission\_source\_id\_refferal .

admission\_source\_id\_transfer -0.241550163

discharge\_Other 0.582398468

change\_No 0.028036698

diabetesMed\_Yes 0.288378150

A1Cresult\_>8 .

A1Cresult\_None 0.227371647

A1Cresult\_Norm 0.004076221

*Table 22: LASSO coefficients of logistic regression model with all predictors*

With an optimal lambda value of approximately 0.0007 chosen through cross-validation, the model effectively balances model complexity and fit, aiding in the selection of the most parsimonious model. Several predictor variables emerged as significant contributors to predicting the outcome. Among these, variables such as *log num lab*, *log num med*, *preceding visits binary*, *log time hosp*, all exhibited positive coefficients. This suggests that an increase in these variables is associated with a higher likelihood of early readmission occurring. Conversely, variable coefficients were negative for categorical variables with levels such as *race = Missing, race = Other, admission source ID = Missing, admission source ID = transfer, log scores* and *A1Cresult = >8*, suggesting that certain subgroups or conditions within these variables is associated with a decreased likelihood of early readmission.

Interestingly, the categorical variable *A1Cresult = Norm* showed a coefficient of only a .0041, very close to 0. This implies that this category does not have a substantial impact on predicting early readmission. The discovery that a normal HbA1c test result during a patient encounter has a smaller effect on the model compared to *AICresult* = *None* adds an intriguing dimension to the understanding of factors influencing early hospital readmission rates. This result echoes the findings of Strack et al. (2014), who observed a similar phenomenon. Their study revealed that regardless of the actual outcome of the HbA1c test, the act of being tested itself was linked to lower readmission rates. As the LASSO coefficient for *A1Cresult = None* is positive, it implies that patients lacking recorded HbA1c values may face a higher predicted likelihood of adverse outcomes.

#### **MODELING**

The model building phase encompasses the selection of suitable models, validation of their assumptions, and adjustment of parameters to enhance predictive accuracy. Since models vary in their assumptions and requirements, data must be tailored accordingly, often necessitating iterative revisits to the data preparation stage. This section of the analysis will examine the application of various machine learning techniques and the adaptations made to address specific model requirements and encountered challenges.

**Logistic Regression Model Preparation**

Logistic regression is a supervised learning technique that aims to predict the probability of an event occurring based on the values of predictor variables. Built upon the principle of probability theory, logistic regression models the relationship between the probability of the binary outcome variable and the predictor variables through the logistic function. This function transforms the linear combination of predictor variables into a probability score between 0 and 1, representing the likelihood of the event happening.

One of the distinguishing features of logistic regression is its ability to provide interpretable results. Logistic regression directly quantifies the impact of each predictor variable on the outcome variable making it well-suited for prediction tasks where understanding the contribution of individual predictors is crucial. Logistic regression estimates the coefficients of the predictor variables that best fit the observed outcomes which indicate the strength and direction of the relationship between each predictor variable and the probability of the outcome. By interpreting these coefficients, analysts can discern which predictor variables have a significant impact on the likelihood of the event.

***Verifying Model Assumptions***

In ensuring the robustness and validity of the logistic regression (LR) model, several key assumptions were carefully examined. These assumptions are tailored to accommodate the binary nature of the outcome variable in logistic regression and include addressing issues of linearity to the logit of the outcome, independence among observations, and low multicollinearity among predictor variables.

One fundamental assumption of logistic regression is that the y values follow a binomial distribution, reflecting the binary nature of the outcome variable. This assumption is intrinsic to the model and does not need to be tested. In contrast to linear regression, which assumes constant variance (homoscedasticity) of the dependent variable (y), logistic regression acknowledges that the variance of y can vary across different levels of the predictor variables.

Similar to linear regression, maintaining independence of observations is another assumption in logistic regression. To uphold this assumption, records in the data set were selected to ensure that each observation corresponds to a unique encounter with a patient. By reducing the dataset to one encounter per patient, the independence of outcomes for different patients is preserved. This is crucial because the likelihood of one patient being readmitted early should not influence the outcome of other patients.

Logistic regression also assumes that the relationship between the predictors and the log odds of the outcome is linear. This assumption implies that the effect of a one-unit change in the predictor variable is constant across all values of that predictor, when holding other predictors constant. To address this assumption, each numeric predictor was individually examined through quantile-odds plots to assess linearity on the logit scale.

A graph with green dots and white text

Description automatically generated

*Figure 26: Number of medications against log odds of readmitted*

To construct the plots, the quartiles were extracted from the summary statistics of each variable. These quartiles are used to divide the data into quantile ranges, and the proportion of observations falling within each range is determined. These proportions represent the probability of an observation belonging to a specific quantile range. Then, the log odds of readmission are computed based on these probabilities using the logit function. The log odds provide a transformed representation of the probability of readmission, enabling a linear relationship to be examined. Finally, the median value of the variable within each quantile range is calculated to represent the central tendency of the data. These median values are plotted against the corresponding log odds of readmission. The resulting plot illustrates the relationship between the selected variable (*x* axis) and the log odds of readmission (*y* axis). See *Figure* *26, Figure 27* and *Figure 28* for these plots.

A graph with green dots and numbers

Description automatically generated

*Figure 27: Time in Hospital against log odds of readmitted*

None of the presented plots exhibit signs indicative of completely linear relationships. However, this lack of strict linearity is not a cause for concern, as the data does not appear entirely random either. While it's observed that in some predictors, values in the third quartile (Q3) surpass the fourth quartile (Q4), the overarching trend remains consistent. Typically, the first plot point (Q1) represents the smallest value, succeeded by Q2, and while Q3 and Q4 may not consistently align perfectly linearly, their progression shows enough signs of linearity to suggest a discernible pattern. See appendix C to view these plots.

A graph with a line and a point

Description automatically generated with medium confidence

*Figure 28: Number of procedures against log odds of readmitted*

***High lever*age *points***

To further ensure the data was fit for logistic regression modeling, high-leverage points were identified. High-leverage points, or values far from the mean in the predictor space, can skew parameter estimates and affect the model's ability to accurately generalize to new data. By identifying and addressing these points, model robustness and validity can be improved.

To assess the impact of high-leverage points in the data, a logistic regression model was fitted and leverage values for each observation were calculated. The logistic regression model includes oversampled positive records, and logarithmic transformations to numeric predictor variables. Leverage points exceeding a specified threshold (calculated as where p represents the number of predictors in the model, n denotes the number of observations), were identified and considered high leverage. Then, subsets of the original dataset were created. These subsets include the data without high leverage values at the specified cutoff values.

The logistic regression model is then refitted using the subsets without high leverage points to assess any impact on model performance. Binary predictions were made for evaluation in a confusion matrix. Finally, the Hosmer-Lemeshow test is employed to evaluate the model's goodness-of-fit, providing insights into how well the model's predictions align with the observed outcomes. Model accuracy, sensitivity, specificity, model AIC and results of the Hosmer-Lemeshow test for each trial are provided in *Table 21*.

Upon excluding high leverage points, the model showed improvements in sensitivity, accuracy, and specificity. This iterative process encompassed evaluating cutoffs at 1.5, 2, 2.5, 2.75, and 3 times the ratio of predictors to observations. Initially, employing a cutoff of 2 led to the exclusion of 465 records, enhancing model sensitivity by 5.415%.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model Metrics: Assessing impact of High-leverage points | | | | | | |
| Cutoff | n | Accuracy | Sensitivity | Specificity | AIC | Hosmer-Lemeshow Test |
| No cutoff | 60,430 | 0.6206 | 0.57255 | 0.62789 | 80,109 | X-squared = 60430,  df = 8,  *p-*value < 2.2e-16 |
|  | 57,383 | 0.6235 | 0.6317 | 0.5691 | 76,560 | X-squared = 57383,  df = 8,  *p-*value < 2.2e-16 |
|  | 59,965 | 0.6197 | 0.6267 | 0.5734 | 80,109 | X-squared = 59965,  df = 8,  *p-*value < 2.2e-16 |
|  | 60,406 | 0.62 | 0.6272 | 0.5726 | 80,695 | X-squared = 60406,  df = 8,  *p-*value < 2.2e-16 |
|  | 60,422 | 0.6202 | 0.6272 | 0.5734 | 80,715 | X-squared = 60422,  df = 8,  *p-*value < 2.2e-16 |
|  | 60,429 | 0.6206 | 0.6279 | 0.5726 | 80,723 | X-squared = 60429,  df = 8,  *p-*value < 2.2e-16 |

*Table 21: Results of LR models of different trials excluding high lever*age *points*

while initially promising, subsequent experimentation revealed that comparable results could be achieved with less data removal. The highest sensitivity occurs with a high-leverage cutoff at 3 times the ratio of predictors to observations. Remarkably, this only necessitated the removal of a single data point – record 568. Consequently, modeling will continue without this solitary outlier.

While it's a common practice to calculate high leverage points using a ratio of 2 times the predictors to observations, experimentation with a ratio of 1.5 revealed the highest performance yet, achieving a sensitivity of 63%. However, this adjustment resulted in the removal of over 3,000 records from the dataset. Considering that the sensitivity improved by just 1%, it raises concerns about the significant loss of valuable information. Therefore, it's prudent to retain this data and explore alternative strategies to enhance performance without sacrificing valuable information.

After preprocessing steps including variable transformations, addressing high-leverage points, validating model assumptions, and tackling data imbalances, a multivariable logistic regression model was constructed. This model utilized readmission as the dependent variable, incorporating predictor variables such as age, race, *gender*, admission type and source IDs, change in medications, HbA1c test results, discharge status, logarithmic transformations of hospitalization time and laboratory tests, and log Charlson scores.

Using the *trainControl* function from the caret package in R, a train control object was set as TC <- trainControl(method = CV, number = 10). This object is used to define the parameters for the training process, such as the resampling method and the number of folds for cross-validation. The number of folds was determined by number = 10, meaning the dataset will be divided into 10 equal-sized folds for training and testing the model. Each fold will be used as a validation set once, while the remaining folds are used for training. This process is repeated 10 times, with each fold serving as the validation set exactly once.

Initially the multivariable logistic regression model was built using unbalanced training data. This model produced no positive predictions, resulting in a sensitivity of 0%, and an accuracy of 86.81%. Here, the high accuracy is likely due to the heavily imbalanced nature of the positive class where the model might have learned to always predict the majority class. For this reason, balancing the data is a crucial step to produce a meaningful model. To assess whether model performance would be affected by eliminating insignificant predictors, coefficients associated with *p-*values exceeding 0.05 were removed from the model. There was little change in the results, with an accuracy of 86.82%, sensitivity of 0.00 and specificity of 1.0 However, the model still made all negative (0) predictions.

**Data Balancing**

To address these class imbalances, oversampling techniques were implemented. First, *readmitted* was balanced by oversampling the positive records with replacement until the number of positive records was equal to those of the negative class, resulting in 30,215 records for both classes. To oversample the positive records, the data was first split into training (80%) and testing (20%) sets using a random selection approach. Performing oversampling on only the training set is crucial to prevent data leakage and maintain the integrity of the testing set. If oversampling were applied to the entire dataset, information from the testing set would be included in the oversampled data, leading to artificially inflated performance metrics during model evaluation.

When using the oversampled data, we see a significant change in accuracy, decreasing to 62.05% and sensitivity which increases to 58.826%. This suggests the model has found unique features of the positive classes and is better able to identify them. Using the oversampled data with all contending predictors, we see the only variable that is not significant in the model is *gender*. All remaining variables exhibit at least one significant level. Upon excluding *gender* from the model, we observe a slight decrease in accuracy to 60.76%. However, as sensitivity remains unchanged, this decline does not warrant the removal of *gender* from the model.

***SMOTE***

To see if the predictive performance of the model could be further improved through other balancing methods, Synthetic Minority Over-sampling Technique (SMOTE) was employed. To mitigate class imbalances, SMOTE generates synthetic examples of the minority class. SMOTE works by identifying instances belonging to the minority class, and for each minority class instance, SMOTE selects its *k*-nearest neighbors in the feature space, where *k* is user-defined.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression Model Metrics | | | | | |
|  | Accuracy | Sensitivity | Specificity | AUC | AIC |
| All predictors,  unbalanced data | 0.8681 | 0.0000000 | 0.9998676 | 0.49993 | 26030 |
| Significant predictors,  unbalanced data | 0.8682 | 0.00000 | 1.0000 | 0.5 | 26063 |
| All predictors,  balanced data | 0.6205 | 0.58862% | 0.62562 | 60.69 | 86,592 |
| Significant predictors,  Balanced data | 0.6264 | 58.46 | 0.6222 | 0.618355 | 79618 |

*Table 22: Logistic regression model metrics: unbalanced vs balanced data*

Next, SMOTE creates synthetic instances by interpolating between the minority class instance and its neighbors. Specifically, it selects one of the *k*-nearest neighbors and generates a new instance along the line segment connecting the two points in the feature space. These synthetic instances are added to the dataset, effectively increasing the representation of the minority class. See figure 29 for a visual representation of how SMOTE addresses class imbalances.

A diagram of different colors of a smote

Description automatically generated with medium confidence

*Figure 29: Visual representation of SMOTE by Soundrapandiyan et al. (2023)*

Implementing SMOTE results in an augmented dataset that combines the original data with the number of synthetically generated data points determined by the user. Variations of SMOTE have been developed to cater to different data distributions and improve classification performance. Among these variations are basic SMOTE, Borderline SMOTE (BL SMOTE), and Adaptive Synthetic Algorithm (ADASYN). Each of these variations were tested.

Basic SMOTE, introduced by Chawla et al. in 2002, randomly selects minority class instances and generates synthetic examples along line segments joining their k nearest neighbors. BL SMOTE, an extension of basic SMOTE, focuses on borderline instances near the decision boundary, thereby reducing the likelihood of generating noisy samples. It achieves this by considering only instances that are close to the decision boundary for synthetic sample generation. Different from basic and BL SMOTE, ADASYN works by adjusting the number of synthetic samples generated for each minority class instance based on the local density of instances and their nearest neighbors. Each SMOTE variation introduces different parameters, such as *k* for controlling the number of nearest neighbors and *dupsize* to fine-tune the sampling process.

In a SMOTE model, *k* denotes the number of nearest neighbors considered during the sampling process which affects the diversity and density of the synthetic samples. A higher value of *k* leads to a smoother interpolation of the feature space and may result in more varied synthetic instances. On the other hand, a lower value of *k* may produce synthetic instances that closely resemble the original minority class instances. *Dupsize indicates the* desired ratio of synthetic minority instances to the original majority instances, with 0 indicating duplication until class balance is achieved.

Given these parameters, lower *k* values were emphasized to best preserve the information within the minority class. Different values of *dupsize* were tested to assess the impact of this parameter, but ultimately, values were chosen that best balance the number of positive and negative records on a 1:1 ratio. This approach ensures that the oversampled dataset maintains a more equitable distribution between the minority and majority classes, facilitating better model training and performance evaluation across both classes.

SMOTE requires numeric input, therefore one-hot encoding was implemented to create a set of dummy variables in both the training and test datasets using the *fastDummies* package. *Dummy* variables serve as numeric representations of categorical variables, where each level of the variable is assigned a new field indicating the presence or absence of the condition through a binary representation. The dummy variables are then used as input for the SMOTE algorithm.

A series of logistic regression models were trained on the SMOTE-tuned datasets using the *glm()* function and evaluated using *k*-fold cross-validation. The seed was set as *123* for all trials. The models used *readmitted* as the outcome variable and *number of medications, log of time in hospital, log of number of lab procedures, preceding visits (binary), log of Charlson scores,* age*, race, gender, admission source ID, admission type ID, change in medication, A1Cresult,* and *discharge status* were used as predictors*.* Confusion matrices were computed to assess model performance, as shown in *Table 23.* Results of the logistic regression model results suggest that increasing the *dupsize* parameter has a notable impact on the performance metrics of the model.

As *dupsize* increases from 2 to 8 in the base SMOTE model, sensitivity consistently improves, indicating that it becomes better at correctly identifying positive instances. However, this improvement in sensitivity comes at the cost of a decrease in overall accuracy. This trade-off suggests that while increasing the synthetic minority instances leads to more positive predictions, it leads to a decrease in the model's overall predictive performance and reliability. This means finding the point where the benefits of oversampling the minority class balance with the potential drawbacks is crucial.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model Metrics: SMOTE | | | | | | |
| *dupsize* | Y = 0 | Y = 1 | Accuracy | Sensitivity | Specificity | Pos. Pred Value |
| 0 | 30,215 | 30,168 | 0.6458 | 0.52358 | 0.66408 | 0.18869 |
| 2 | 30,215 | 13,776 | 0.8388 | 0.12063 | 0.94747 | 0.25794 |
| 4 | 30,215 | 22,960 | 0.7208 | 0.44231 | 0.76300 | 0.22029 |
| 5 | 30,215 | 30,168 | 0.6447 | 0.52114 | 0.66311 | 0.18754 |
| 6 | 30,215 | 32,144 | 0.5894 | 0.63636 | 0.58224 | 0.18739 |
| 7 | 30,215 | 36,736 | 0.5294 | 0.71329 | 0.50152 | 0.17805 |
| 8 | 30,215 | 41,328 | 0.4721 | 0.7736 | 0.4265 | 0.1696 |

*Table 23: SMOTE model metrics*

In the end it was decided of the base SMOTE models, using *dupsize = 0 (or dupsize* = 6) produces a model with the best results. Using these sizes ensures the number of positive and negative records is as close as possible. Of the models with an accuracy of at least 60%, the highest sensitivity observed is 52.114%, using *dupsize* = 5. Although the model using *dupsize* = 6 does not achieve the goal of at least 60% accuracy, it does produce the most balanced results of an accuracy of 58.94%, sensitivity of 63.36% and specificity of 58.244%.

***ADASYN***

ADASYN (Adaptive Synthetic Sampling) introduces several features that distinguish it from other oversampling techniques due to its ability to adapt its sampling strategy based on the data distribution. ADASYN focuses on synthesizing more samples for instances that are harder to classify, effectively addressing regions of the feature space where the minority class is underrepresented. Specifically, it achieves this by iteratively generating synthetic samples for minority class instances that are located in densely populated regions of the feature space, where the density of the majority class is higher. This adaptability enables ADASYN to prioritize the creation of synthetic samples near the decision boundary, enhancing the model's ability to learn from challenging instances and improve classification performance.

In ADASYN (Adaptive Synthetic Sampling), the parameter *k* determines the number of nearest neighbors used to generate synthetic data points for in the minority class. A higher value of *k* means that more neighbors are considered when generating synthetic samples, which can lead to a smoother interpolation in the feature space. This parameter directly impacts the diversity and density of the synthetic samples created. See *Table 24* for results of the ADASYN rebalancing.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model Metrics: ADASYN | | | | | | |
| *k* | Y = 0 | Y = 1 | Accuracy | Sensitivity | Specificity | Pos. Pred Value |
| 2 | 30,215 | 32,737 | 59.35% | 59.756% | 59.287 | 17.76% |
| 3 | 30,215 | 31,912 | 60.04% | 58.049 | 60.342 | 17.926 |
| 5 | 30,215 | 32,882 | 59.65 | 58.699 | 59.797 | 17.889 |
| 10 | 30215 | 31,609 | 62.16% | 55.610% | 63.132% | 18.372% |
| 15 | 30215 | 32,737 | 60.04 | 58.049 | 60.432 | 17.926% |
| 20 | 30,215 | 32,470 | 60.43% | 57.886% | 60.815% | 18.062% |
| 50 | 30,215 | 32,481 | 60.28 | 58.374 | 60.560 | 18.090 |
| 100 | 30,215 | 32,979 | 60.31 | 58.211 | 60.621 | 18.072 |
| 200 | 30,215 | 32,880 | 60.79 | 57.805 | 61.240 | 18.203 |

*Table 24: ADASYN Model results*

The process of finding the optimal tuning parameters for the ADASYN model entails testing values of *k* in an iterative, stepwise fashion. When testing different values of *k*, the tradeoff between sensitivity and specificity was minimal, as *k* increases*.* Both specificity and sensitivity hovered around 60% across all models. While employing higher *k* values seems to increase the number of positive predictions, among models with over 60% accuracy, none achieve greater than 62.8% sensitivity.

Tuning *k* aids in improving performance, however relying on a specific *k* value or seed for desired outcomes is unreliable due to the stochastic nature of the algorithm. The analysis reveals that tuning the ADASYN model has minimal impact on performance, regardless of the value of *k*, but it does contribute to a slight increase in the number of predictive positives. However, the most significant factor influencing performance is the size of the synthetic positive records. Upon examining the metrics, we observe that models with higher accuracy tend to have a greater number of synthetic positive records. It's noteworthy that the ADASYN model lacks a parameter to control the quantity of synthetic samples generated; instead, it relies on the seed and *k* size. Consequently, while the ADASYN model may perform well in certain trials, we cannot solely rely on randomness to consistently produce favorable results.

***Borderline SMOTE***

Unlike traditional SMOTE, which synthesizes new instances indiscriminately across the feature space, Borderline-SMOTE (BL SMOTE) focuses on the borderline instances - those minority class instances that are near the decision boundary between the minority and majority classes. By specifically targeting these borderline instances, BL SMOTE aims to address the regions where classification is most challenging, thereby improving the model's ability to discriminate between classes. This targeted oversampling strategy helps mitigate the risk of oversampling noise or introducing synthetic samples in regions already well-represented by the minority class.

Similar to ADSYSN, in BL SMOTE, *k* denotes the number of nearest neighbors considered during the sampling process. *C* represents the number of nearest neighbors utilized during the calculation of safe-levels, which is a measure of confidence in the classification of minority instances calculated based on the distribution of majority and minority instances around a given data point. *Dupsize* indicates the desired ratio of synthetic minority instances to the original majority instances, with 0 indicating duplication until class balance is achieved.

Results of the BL SMOTE model highlight a dramatic tradeoff between sensitivity and specificity. Different values of *C* were tested, which appears to have a positive relationship with model sensitivity, at the costs of lowered accuracy. Similar to ADASYN, increasing k results in lowered sensitivity due to a dilution effect caused by oversampling a larger number of minority class instances. This dilution effect occurs because as the number of synthetic minority class samples increases, the influence of each individual minority instance on the decision boundaries of the classifier decreases, leading to decreased sensitivity.

The BL SMOTE model that produces the optimal tradeoff point for these metrics is when *dupsize* = 13 (or *dupsize* = 0). The balance between sensitivity (61.1%) and specificity (60.580%) at *dupsize* = 13 suggests that the model neither overly emphasizes sensitivity at the expense of accuracy nor prioritizes accuracy over sensitivity.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model Metrics: Border Line SMOTE | | | | | | | |
| *dupsize* | *k* | Y = 0 | Y = 1 | Accuracy | Sensitivity | Specificity | Pos. Pred Value |
| 0 | 2 | 30215 | 30507 | 0.6317 | 0.5407 | 0.6453 | 0.1853 |
| 5 | 2 | 30215 | 14092 | 0.8612 | 0.045455 | 0.984650 | 0.309524 |
| 10 | 2 | 30215 | 27392 | 0.7498 | 0.35927 | 0.80892 | 0.22156 |
| 12 | 2 | 30215 | 27392 | 0.6569 | 0.54720 | 0.67355 | 0.20239 |
| **13** | **2** | **30215** | **29292** | **0.6064** | **0.61014** | **0.60580** | **0.18983** |
| 14 | 2 | 30215 | 31192 | 0.5522 | 0.67220 | 0.53407 | 0.17925 |
| 15 | 2 | 30215 | 33092 | 0.4964 | 0.73427 | 0.46037 | 0.17080 |
| 20 | 2 | 30215 | 42592 | 0.2793 | 0.9240 | 0.1817 | 0.1460 |
| 50 | 2 | 30215 | 99592 | 0.1325 | 0.999126 | 0.001323 | 0.131530 |

*Table 25: Borderline SMOTE model metrics*

***Results of Data Balancing***

The unbalanced data set achieves the highest accuracy among all the sampling techniques, yet its sensitivity is 0.00, indicating a failure to identify any positive instances. In contrast, the balanced data set exhibits a more equitable distribution across accuracy (62.26%), sensitivity (61.54%), and specificity (62.37%) and positive predictive value, making it a more reliable choice. While SMOTE techniques attempt to address class imbalance, the BL SMOTE and base SMOTE fall short in achieving comparable performance to the balanced data set across multiple metrics.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Final SMOTE Model Metrics | | | | |
|  | Accuracy | Sensitivity | Specificity | Pos. Pred Value |
| SMOTE base | 64.58 | 52.358% | 66.408% | 18.869% |
| ADASYN | 59.35% | 59.756% | 59.287% | 17.76% |
| SMOTE BL | 60.64% | 61.014% | 60.580% | 18.983% |
| Balanced Data | 62.26% | 61.538% | 62.366% | 18.423% |
| Unbalanced Data | 0.8682 | 0.00000 | 1.0000 | 0.00 |

*Table 26: Final logistic regression model metrics*

While SMOTE-based methods did not outperform the straightforward oversampling approach, it was still worth exploring. Even marginal improvements in effectiveness can translate to significant benefits in real-world applications, particularly in domains where class imbalance poses critical challenges, such as medical diagnosis. Additionally, investigating SMOTE techniques provides insights into the comparative advantages and limitations of different approaches to addressing class imbalance, enhancing the understanding of their underlying mechanisms. Thus, despite the modest performance gain, ultimately the unbalanced data proved to be the most reliable and effective method.

**Final Logistic Regression Model**

The final multivariable logistic regression was built using the balanced data and utilizes *readmission* as the dependent variable, with predictor variables age*, race, gender, admission type and source IDs, change in medications, HbA1c test results, discharge status,* the log of *hospitalization time, laboratory test*s, and *Charlson scores*. K-fold cross validation was implemented to assess potential overfitting. Model output can be seen in *Table 27.*

Assessing logistic regression model summaries contribute to a comprehensive understanding of the predictors' effect on the outcome variable. The estimate is formed by iteratively adjusting coefficients to maximize the likelihood of observing the given outcomes, creating the best-fitting relationship between predictor variables and the log-odds of the outcome. Standard errors quantify the variability of coefficient estimates, offering insights into the precision of the model's parameter estimates. Meanwhile, *t*-values and corresponding *p-*values assess the statistical significance of each coefficient, indicating whether the observed effects are likely to be due to chance.

*Table 27: Final Logistic Regression (SMOTE) Model Summary*

Call:

glm(formula = readmitted\_2 ~ ., family = binomial, data = data\_subset)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.7984568 0.0835287 -21.531 < 2e-16 \*\*\*

raceCaucasian 0.0778818 0.0212704 3.662 0.000251 \*\*\*

raceMissing -0.2048932 0.0525689 -3.898 9.71e-05 \*\*\*

raceOther -0.1976759 0.0431851 -4.577 4.71e-06 \*\*\*

genderMale -0.0139739 0.0160773 -0.869 0.384753

admission\_type\_idElective -0.2140340 0.0399569 -5.357 8.48e-08 \*\*\*

admission\_type\_idEmergency\_or\_trauma -0.2321442 0.0370807 -6.261 3.84e-10 \*\*\*

admission\_type\_idUrgent -0.0965799 0.0381914 -2.529 0.011444 \*

admission\_source\_idER 0.2468480 0.0454505 5.431 5.60e-08 \*\*\*

admission\_source\_idOther 0.2619978 0.0452776 5.786 7.19e-09 \*\*\*

admission\_source\_idrefferal 0.0643370 0.0441350 1.458 0.144914

admission\_source\_idtransfer -0.1333574 0.0515352 -2.588 0.009662 \*\*

log\_num\_lab 0.1086691 0.0114863 9.461 < 2e-16 \*\*\*

log\_num\_med 0.0680777 0.0197028 3.455 0.000550 \*\*\*

preceding\_visits\_binary 0.7984369 0.0246175 32.434 < 2e-16 \*\*\*

changeNo 0.0488558 0.0189204 2.582 0.009818 \*\*

diabetesMedYes 0.2865033 0.0221347 12.944 < 2e-16 \*\*\*

A1Cresult>8 0.0475206 0.0490637 0.969 0.332770

A1CresultNone 0.3052239 0.0416196 7.334 2.24e-13 \*\*\*

A1CresultNorm 0.0740764 0.0529057 1.400 0.161466

log\_time\_hosp 0.3893001 0.0174682 22.286 < 2e-16 \*\*\*

log\_scores -0.0008895 0.0155577 -0.057 0.954407

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 91121 on 65729 degrees of freedom

Residual deviance: 88111 on 65708 degrees of freedom

AIC: 88155

Number of Fisher Scoring iterations: 4

Lower *p-*values and larger absolute *t*-values suggest stronger evidence against the null hypothesis, supporting the confidence in the significance of the predictors. Confidence intervals provide a range of plausible values for each coefficient, accounting for uncertainty in the estimation process. By considering these additional outputs alongside coefficient estimates, the significance and reliability of each predictor's effect on the outcome variable can be interpreted.

Assessing variable coefficients from the model summary we see *hospitalization duration*, *number of laboratory tests,* and *preceding visits* exhibit positive coefficients, indicating a positive association with readmission likelihood. Conversely, higher medication counts, and certain racial categories demonstrate negative coefficients, suggesting a protective effect against readmission. Demographic and clinical factors like being of older age, or a male and higher medication use also play pivotal roles.

**Odds Ratios**

Comparing the exponentiated coefficients of the logistic regression model provides valuable insights in the form of an odds ratio. The odds ratio (OR) is a measure of association between a predictor and an outcome, quantifying the strength and direction of the association between the exposure and the outcome. The odds ratio takes on the form:

Odds Ratio =

where each letter represents a specific count in a 2x2 contingency table. The letter *a* denotes the count of exposed individuals who experience the outcome of interest, while *b* represents the count of exposed individuals who do not experience the outcome. Similarly, *b* signifies the count of unexposed individuals who experience the outcome, and *d* represents the count of unexposed individuals who do not experience the outcome.

When the odds ratio surpasses 1, it signifies an increased likelihood of the outcome associated with the exposure variable, suggesting individuals exposed are more prone to experience it. Conversely, when the odds ratio falls below 1, it denotes a decreased likelihood of the outcome linked with the exposure variable. Comparing odds ratios is paramount as it reveals significant patterns and directions of predictive power. Despite the structural limitations of this data, even with variable standardization, this comparative approach allows prioritization of predictors based on their relative impact, rather than solely on individual odds ratios.

The odds ratio results reveal insights into the association of categorical variables by quantifying how the odds of the outcome differ between different categories of the variable. The most impactful variable in the final model appears to be *preceding visits,* with an odds ratio of 2.22 indicates that patients who visited either the ER, inpatient or outpatient services within the previous year are more than twice as likely to be readmitted compared to those without prior visits. This highlights the importance of continuity of care and proactive management for patients with a history of hospitalizations.

Odds Ratios

raceCaucasian 1.08099

raceMissing 0.81473

raceOther 0.82063

genderMale 0.98612

admission\_type\_idElective 0.80732

admission\_type\_idER\_trauma 0.79283

admission\_type\_idUrgent 0.90793

admission\_source\_idER 1.27998

admission\_source\_idOther 1.29952

admission\_source\_idrefferal 1.06645

admission\_source\_idtransfer 0.87515

log\_num\_lab 1.11479

log\_num\_med 1.07044

**preceding\_visits\_binary 2.22206**

changeNo 1.05006

diabetesMedYes 1.33176

A1Cresult>8 1.04866

A1CresultNone 1.35692

A1CresultNorm 1.07688

log\_time\_hosp 1.47594

log\_scores 0.99911

*Table 28: Odds ratios of final logistic regression model*

The Charlson Comorbidity Index (CCI) measures the complexity of a patient's health conditions, with higher scores indicating more serious comorbidities. With an odds ratio of 0.99, the data shows that as CCI scores increase, the likelihood of readmission decreases by about 1% for each unit rise in the logarithm of CCI scores. This suggests that there is minimal change in the probability of readmission for each unit rise in the logarithm of CCI scores. Despite establishing a significant association between CCI scores and readmission status in earlier EDA steps, it’s possible the Charlson Comorbidity Index (CCI) may not capture the specific health conditions or severity levels that strongly influence readmission risk in the studied population, leading to an underestimation of the true impact of comorbidities on readmission risk in the studied population.

The odds ratio of 1.07 for the log of the number of medications a patient is prescribed suggests that with each additional unit increase in the logarithm of *num medications*, there's a 7% rise in the odds of readmission. This signifies that as patients receive a greater number of medications, their likelihood of being readmitted rises. This association may be indicative of the complex health conditions or multiple comorbidities often presented in patients requiring numerous medications. These individuals are typically managing severe or chronic illnesses, making them more vulnerable to readmission due to the heightened complexity and severity of their health issues.

The *log of time in hosp*ital odds ratio of 1.47 indicates that for every unit increase in the logarithm of total days in the hospital, the odds of readmission increase by approximately 47%. This suggests that prolonged hospital stays are associated with a higher risk of readmission. Prolonged hospital stays often indicate more severe or complex health conditions requiring extensive treatment and management. Patients with such conditions may be at a higher risk of experiencing complications during their hospitalization, increasing the likelihood of readmission after discharge.

Similarly, the *log num lab* odds ratio of 1.1147 suggests that for each additional unit in the logarithm of the number of lab procedures performed, the odds of readmission increase by approximately 11.5%. The frequency of lab procedures may reflect the need for ongoing disease monitoring, treatment adjustments, or response assessments. Patients requiring more frequent testing may have conditions that necessitate closer medical oversight and intervention, indicating a higher risk of readmission if their condition worsens or requires further management.

*Admission type* *ID* classifies the urgency and nature of how a patient is admitted (elective, urgent, ER, missing). The reference level for *admission type* *ID* is set to *missing* with an odds ratio of 1. Patients admitted for elective procedures exhibit an odds ratio of 0.807, indicating approximately an 80.7% lower odds of readmission compared to those with missing admission types, when controlling for other variables in the model. Similarly, patients admitted for emergency or trauma cases demonstrate an odds ratio of 0.793, suggesting approximately a 79.3% lower odds of readmission relative to those with missing admission types. Furthermore, patients admitted for urgent care display an odds ratio of 0.908, indicating approximately a 90.8% lower odds of readmission compared to the reference category.

By setting the reference level for *admission type ID* to *missing*, the impact of various admission types on readmission likelihood relative to those with missing admission data could be assessed. Notably, all other odds ratios for admission types were less than 1, indicating a lower likelihood of readmission compared to patients with missing admission data. This suggests that having missing data is actually the most associated factor with an early hospital readmission. This finding underscores the importance of comprehensive data collection and documentation practices.

*Admission source ID* indicates the origin from which a patient enters in a hospital stay. Patients admitted from the emergency room exhibit an odds ratio of 1.28, suggesting approximately a 28% higher likelihood of readmission compared to those with missing admission data. Similarly, patients admitted from other sources besides ER, referral, or transfer demonstrate an odds ratio of 1.299, indicating around a 30% higher likelihood of readmission relative to those with missing admission data. Conversely, patients referred from other healthcare facilities exhibit an odds ratio of 1.066, indicating a modest 6.6% higher likelihood of readmission compared to the reference category. Interestingly, patients transferred from another healthcare facility display the lowest odds ratio of 0.875, implying approximately an 12.5% lower likelihood of readmission compared to those with missing admission data.

Patients with HbA1c levels greater than 8, as indicated by an odds ratio of 1.048 compared to HbA1c > 7, have poorly controlled blood sugar levels, signifying uncontrolled diabetes. This subgroup faces a 4.8% increased likelihood of readmission. Patients with no recorded HbA1c result demonstrate an odds ratio of 1.3569 compared to HbA1c> 7, implying a significant 35.7% increase in readmission likelihood. This absence of HbA1ctesting suggests incomplete documentation or missed opportunities for diabetes monitoring. Without proper monitoring and management of blood sugar levels, these patients face heightened risks of acute complications and hospital admissions, resulting in elevated readmission rates.

Patients with normal HbA1cresults within the target range, as indicated by an odds ratio of 1.0768, suggest well-controlled blood sugar levels and effective diabetes management. Interestingly, this odds ratio, being slightly above 1, indicates that a normal result is more impactful on early hospital readmission than having a high score (>8 or >7). The inference is that normal HbA1c scores may not notably influence readmission rates. Normal test results affirm that the patient's diabetes is well-regulated, with blood sugar levels within the target range. Consequently, these patients may not necessitate additional services that would typically be required for those with extremely high levels. These services may serve as a preventative measure or buffer between another early admission, helping to maintain stability in the patient's condition and mitigate the likelihood of readmission.

**Further Exploration of HbA1c**

Strack et al (2014) initially investigated the relationship between HbA1c testing and primary diagnoses across various patient populations. They observed that patients with a primary diagnosis of diabetes were more frequently tested for HbA1c compared to those diagnosed with respiratory or circulatory conditions. Seeking further insight, the focus shifted to exploring potential distinctions between patients with a primary diagnosis of uncontrolled diabetes, controlled diabetes, and unspecified diabetes with the goal of understanding whether differences in admission profiles exist among these patient groups. Utilizing the final logistic regression model, patient subsets were delineated for each category to evaluate any discernible dissimilarities in patient profiles

The models were built using the balanced data with predictor variables age*, race, gender, admission type and source IDs, change in medications, HbA1c test results, discharge status,* logarithmic transformations of *time in hospital, laboratory tests,* and *Charlson scores*. See *Table 29* for comparison of Hba1C logistic regression results among different patient subgroups. The reference level for *A1Cresult* is set to *>7* indicating that it serves as the baseline category for comparison.

***Unspecified Diabetes***

Patients with unspecified diabetes and normal HbA1c results demonstrated no change in the odds of early readmission compared to those with HbA1c levels greater than 7, with an odds ratio of 1.0. Similarly, patients with no HbA1c result exhibited an odds ratio of 0.997, suggesting that not having a HbA1c test does not increase the odds of readmission among this group. However, among patients with unspecified diabetes and an HbA1c test result of *>8*, exhibit an odds ratio of 0.534, suggesting that these patients are nearly half as likely to face an early readmission compared to results *>7.*

Among all 4 levels of *A1Cresult*, the significance of only *A1Cresult = 8* may be indicative of interventions or treatment strategies specifically targeting patients with higher HbA1c levels. A HbA1c result of 7 is not as extreme as 8, suggesting that patients with HbA1c levels hovering around 7 may not receive as much intensive care or attention as those with higher scores.

***Uncontrolled diabetes***

For patients with a primary diagnosis of uncontrolled diabetes, the logistic regression model produced no significant coefficients for *A1Cresult > 8* or *A1Cresult* = *missing*. However, the model revealed that individuals with uncontrolled diabetes and a normal HbA1c result were significantly less likely to face an early hospital readmission compared to those with HbA1c levels greater than 7, with an odds ratio of 0.2861 and a *p-*value of 0.016. This result suggests that achieving a normal HbA1c result for patients with uncontrolled diabetes may be associated with improved disease management or less severe disease status, leading to decreased odds of early readmission.

*Table 29: Logistic regression model summaries by primary diagnosis subsets*

Estimate Std. Error z value Pr(>|z|) Odds Ratio

**All Diabetes**

A1Cresult\_>8 -0.767469 0.192238 -3.992 6.54e-05 \*\*\* 0.464

A1Cresult\_None -0.054624 0.180499 -0.303 0.762174 0.9468

A1Cresult\_Norm -0.176010 0.228124 -0.772 0.440380 0.838

**Uncontrolled Diabetes**

A1Cresult\_>8 -0.7516069 0.4194063 -1.792 0.073121 . 0.471

A1Cresult\_None -0.1022799 0.4128621 -0.248 0.804340 0.902

A1Cresult\_Norm -1.2511755 0.5217345 -2.398 0.016480 \*\* 0.2861

**Controlled Diabetes**

A1Cresult\_>8 -1.12679 0.84606 -1.332 0.18292 0.3240

A1Cresult\_None 0.39730 0.83252 0.477 0.63320 1.487

A1Cresult\_Norm 1.53339 0.98667 1.554 0.12016 4.6338

**Unspecified Diabetes**

A1Cresult\_>8 -6.273e-01 2.570e-01 -2.441 0.014655 \* 0.534

A1Cresult\_None -3.070e-03 2.251e-01 -0.014 0.989117 0.997

A1Cresult\_Norm 7.343e-05 2.916e-01 0.000 0.999799 1.00

**Circulatory Diagnoses**

A1Cresult\_>8 0.953162 0.242269 3.934 8.34e-05 \*\*\* 2.5938

A1Cresult\_None 1.003243 0.207047 4.845 1.26e-06 \*\*\* 2.727

A1Cresult\_Norm 0.872112 0.249548 3.495 0.000474 \*\*\* 2.392

**Respiratory Diagnoses**

A1Cresult\_>8 0.231703 0.308493 0.751 0.452604 1.206

A1Cresult\_None 0.517553 0.255711 2.024 0.042973 \* 1.6779

A1Cresult\_Norm -0.690004 0.351711 -1.962 0.049780 \* 0.5015

**All Primary Diagnoses**

A1Cresult\_>8 0.186826 0.055959 3.339 0.000842 \*\*\* 1.205

A1Cresult\_None 0.535419 0.047486 11.275 < 2e-16 \*\*\* 1.708

A1Cresult\_Norm 0.031636 0.060847 0.520 0.603115 1.032

While the coefficient for patients who did not receive HbA1c testing is not statistically significant, the odds ratio of 0.902 suggests that these patients are approximately 10% less likely to be readmitted compared to those with HbA1c results greater than 7. Despite lacking statistical significance, this finding aligns with trends of previous conclusions from the analysis, indicating that not receiving any HbA1c testing contributes more significantly to readmission risk than other testing outcomes.

***Circulatory***

In the circulatory patient subset, significant odds ratios emerged for all HbA1c result categories. Patients with HbA1c levels greater than 8 exhibited approximately 2.594 times higher odds of certain outcomes compared to the reference category, indicating a potential link between elevated HbA1c levels and increased likelihood of certain admission profiles. Similarly, patients with no available HbA1c result or a normal HbA1c result also demonstrated significantly increased odds of certain outcomes, with odds ratios of approximately 2.727 and 2.392, respectively. These findings suggest that both high and normal HbA1c results may play a role in predicting certain admission profiles among patients with circulatory conditions.

These results highlight an interesting finding related to the impact of diabetes on vascular health and the associated circulatory conditions. Diabetes is known to significantly impact vascular health, often leading to circulatory issues such as peripheral arterial disease, coronary artery disease and cerebrovascular disease. Elevated HbA1c levels reflect poor glycemic control, which in turn exacerbates vascular complications in diabetic patients. Therefore, the high odds ratios associated with HbA1c results greater than 8 among diabetic patients with circulatory issues may indicate an increased risk of complications or exacerbation of existing circulatory conditions, potentially necessitating more frequent hospital admissions or interventions.

***Respiratory***

Conversely, the analysis of respiratory patients revealed a more nuanced relationship between HbA1c testing results and admission profiles. While patients with HbA1c levels greater than 8 showed a slight increase in odds of certain outcomes (odds ratio of approximately 1.206), this result was not statistically significant. This suggests that among respiratory patients, elevated HbA1c levels may not be as directly linked to early readmission as in circulatory patients, indicating a potentially lesser impact of glycemic control on healthcare utilization in this population.

However, patients with a primary diagnosis of respiratory-related issues and no available HbA1c result exhibited significantly higher odds of early readmission. With an odds ratio of approximately 1.6789, this suggests that the absence of HbA1c testing is associated with increased risk of early readmission, further supporting previous conclusions. Interestingly, patients with a normal HbA1c result demonstrated significantly lower odds of certain outcomes, with an odds ratio of approximately 0.5015, indicating a potential protective effect of normal HbA1c levels against early readmission among respiratory patients.

***All diagnoses***

When interpreting results of HbA1c coefficients of the model containing records of all diabetic patients, not just those with the specified primary diagnoses, the patterns observed in the analysis of these subsets come to light. Interestingly, we see HbA1c results with normal scores do not have a significant effect on readmission outcomes, with an odds ratio of 1.032. This supports previous conclusions that a test alone is not protective of early readmission. However, a significant coefficient for HbA1c = none (< 2e-16 \*\*\*) and odds ratio of 1.708, suggests that while having a test alone may not be protective, having no test performed does increase the odds of early readmission. This result may be attributed not to the results of the test itself, but rather to the responses elicited from healthcare teams.

Despite inconsistencies in the significance of coefficients among different subsets, this comparative analysis highlights previous trends regarding the impact of HbA1c testing. From these analyses, it becomes evident that patients with uncontrolled diabetes benefit most from HbA1c testing, with an odds ratio of 0.286, this result stands out as the most distinct from the baseline level and represents the lowest risk among all possible HbA1c outcomes. These patients exhibit approximately a 71.4% reduction in the likelihood of experiencing early readmission compared to those with test scores exceeding 7.

When comparing subsets of patients with primary diagnoses of uncontrolled, controlled or unspecified diabetes, we observe varied results. This variance may stem from differences in dataset sizes, as well as potential variations in the distribution of patient characteristics or healthcare practices across the subsets. Additionally, with smaller dataset sizes, there is a greater chance that unique patient characteristics may overshadow broader trends, leading to increased variability in the observed result.

Despite the challenges posed by differences in data subset sizes and potential variations in patient characteristics, the aggregation of these results into the final logistic regression model, incorporating all diagnoses, underscores important trends. While the variability in outcomes may require careful interpretation and further investigation, it also offers opportunities for additional analysis and exploration.

***Homogeneity test of proportions***

The difference in the distribution of HbA1c levels between patients across various primary diagnoses was investigated. This was done using a 2x4 matrix representing the counts of HbA1c levels (>8, Norm, >7) for each group and employing a chi square test and Fisher’s exact test to test the homogeneity of proportions among these groups. Groups that were compared were: uncontrolled vs. uncontrolled diabetes, uncontrolled vs. unspecified diabetes, and controlled vs. unspecified diabetes.

In the analysis comparing uncontrolled versus controlled diabetes, the initial examination using all four levels of HbA1c did not yield significant results. Subsequently, records with HbA1c = *None* were removed, leading to a slight improvement in the p-value to 0.5679, but still no significant findings emerged. HbA1c was then converted into a binary variable, distinguishing between *None* and the combined categories of *>7, >8,* and *Normal*. Although this adjustment marginally improved significance, with a p-value of 0.4698 from Fisher's exact test, the results remained statistically insignificant. Therefore, it is evident that there is no significant difference in the distribution of HbA1c levels between uncontrolled and controlled diabetes groups in this analysis.

To test the difference in distribution of Hba1c levels between uncontrolled and unspecified diabetes, initially a chi-squared test was employed. However, due to small expected cell counts among some distributions, a warning was issued regarding the potential inaccuracy of the chi-squared approximation. Nevertheless, the test yielded a p-value of 0.008576, indicating significant evidence to reject the null hypothesis. To validate our findings and address concerns about accuracy, Fisher's exact test was also conducted, returning a p-value of 0.007197, and reaffirming significant evidence against the null hypothesis. Consequently, it was concluded that there is a statistically significant difference in the distribution of HbA1c levels between patients with uncontrolled and unspecified diabetes.

**XGBoost**

XGBoost, created by Chen & Guestrin (2016) is a powerful boosting ensemble method widely employed in regression and classification tasks. It utilizes a gradient boosting framework, unlike traditional decision tree algorithms, to systematically integrate new trees into the ensemble. This approach enables XGBoost to effectively address errors made by existing trees, leading to improved model performance.

XGBoost incorporates an array of regularization techniques, including shrinkage and tree pruning, to combat overfitting and enhance generalization capabilities. This ensures that the model is not overly sensitive to the training data. Furthermore, it leverages features like parallel computing and hardware optimization to expedite training procedures, rendering it particularly adept at handling large-scale datasets. The xgboost model works by:

* Starting with an initial prediction, the mode (most frequent class label)
* Calculating the gradients (errors which indicate the direction and magnitude of the error for each data point) and second-order derivatives (representing the curvature of the loss function with respect to the predictions)
* Training decision trees predict the gradients, added sequentially, each correcting the errors of the previous ones.
* Updating the model's predictions by adding the predictions of the decision trees, scaled by the learning rate (eta)
* Applying regularization techniques to prevent overfitting

The iterative process in XGBoost continues until a stopping criterion is met or until no further improvements are observed. At each iteration, a new decision tree is added to the ensemble to refine predictions. Finally, the model's prediction is calculated as the sum of predictions from all the decision trees in the ensemble, providing a comprehensive and refined estimation.

One important hyperparameter is the tree depth (*max\_depth*), controlled to mitigate overfitting by adjusting the minimum number of samples required to split a node. The default value for *max\_depth* parameter is 6, where higher values, indicate larger tree depths and more complex models. However, the trade-off using higher *max\_depth* values is it increase the risk of overfitting. *Max\_depth* values of 2, 4 and 6 were tested, but remained at 2 to avoid potential overfitting.

Additionally, the learning rate (*eta*) determines the contribution of each tree to the final prediction in the ensemble, where smaller learning rate results in slower learning. The default value for this parameter is 0.3, with possible values ranging 0-1. In essence, the learning rate parameter influences the trade-off between model complexity and generalization ability by shrinking feature rates till they reach the optimal value, while supported by *nrounds,* or the maximum number of iterations. Balancing these hyperparameters requires careful evaluation techniques to determine the most effective combination for achieving the desired model outcomes.

XGBoost was also employed in this analysisto predict early hospital readmissions and better understand variable importance. To prepare the data, training and test sets were constructed on an 80/20 split. Given XGBoost's requirement for numerical inputs, categorical variables were transformed into dummy variables using one-hot encoding with the *dummy* package. Subsequently, DMatrix objects were created to encapsulate the predictor variables and corresponding labels for both the training and test datasets. Parameterization of the XGBoost model was then explored, testing key parameters such as maximum tree depth (*max\_depth*), learning rate (*eta*), and objective function (binary:logistic). Furthermore, the evaluation metric was set to AUC offering insights into the model's performance.

The XGBoost model was trained using the xgb.train() function. The incorporation of the watchlist parameter enabled real-time monitoring of model performance on both the training and evaluation datasets throughout the training process. After training the model, predictions were made on the test dataset. Predicted probabilities were obtained for each observation, and a threshold of 0.5 is applied to convert these probabilities into binary predictions. Finally, model performance was evaluated with a confusion matrix. See *Table 30* for a table of the contending XGBoost models and model metrics.

Initially, XGBoost was trained using unbalanced data, yielding poor results with an overall accuracy of 12.99% and sensitivity of 0.0003%, marking it as the poorest performing baseline model thus far. The number of rounds was set to 10, max tree depth was set to 2, and eta was set to 0.1. These lackluster results are likely due to the imbalanced nature of the outcome variable, making it challenging for the model to identify the positive cases. To address this issue with goals of improving the predictive power of the model, data balancing and tuning methods were implemented.

Subsequently, the SMOTE data was utilized as input for the model. Using the SMOTE data, without any additional SMOTE tuning, accuracy increases from 13.01 to 13.77 and sensitivity increases from 0.0004 to 0.014 compared to the unbalanced data. While these results suggest over-resampling assisted in the identification of positive records, it was not by much and the model requires significant improvement in model performance.

Adjustments were made to various parameters in the XGBoost model. One key parameter that was examined was the number of boosting rounds (*nrounds*). Experimentation was conducted with different values of *nrounds*, including 10, 20 50 and 100. The goal was to find an optimal balance between model complexity and generalization by determining the point at which increasing the number of rounds no longer improved performance. It was observed that increasing *nrounds* beyond 100 had little effect on model performance, suggesting diminishing returns.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| XGBoost Model Performance | | | | | | | |
| Data | *Max\_depth* | eta | *nrounds* | Accuracy | Sensitivity | Specificity | AUC |
| Unbalanced data | 5 | 1 | 10 | 13.01 | 0.0003639 | 0.9991870 | 0.4998 |
| SMOTE | 5 | 1 | 10 | 13.77 | 0.01383 | 0.96748 | 0.5063 |
| SMOTE | 2 | 1 | 10 | 0.2356 | 0.1658 | 0.7033 | 0.5083 |
| SMOTE | 2 | 0.1 | 10 | 33.79 | 30.94 | 52.93 | 0.4193 |
| SMOTE | 2 | 0.1 | 100 | 18.64 | 0.08892 | 0.83984 | 0.4644 |
| SMOTE | 2 | 0.01 | 10 | 40.07 | 0.40034 | 0.40325 | 0.5982 |
| SMOTE | 2 | 0.001 | 10 | 0.4007 | 0.40325 | 0.40325 | 0.5982 |
| SMOTE | 2 | 0.001 | 50 | 0.4007 | 0.40034 | 0.40325 | 0.5982 |
| SMOTE | 3 | 0.001 | 50 | 0.2885 | 0.2335 | 0.6569 | 0.4452 |

*Table 30: Results of XGBoost model tuning*

Concurrently, the learning rate (*eta*) was adjusted. Reducing *eta* from 0.1 to 0.01 improved model AUC by 3.51% and increased model accuracy from 23.56% to 33.79%. Sensitivity also increased from 16.58% to 33.79%. However, when the learning rate is smaller than 0.001, accuracy and sensitivity decline. It was concluded that the optimal value of *eta* is somewhere between 0.1 and 0.01. Various values were tested, and 0.03 was determined as the optimal value. Following iterative trial and error of model tuning, the parameters used in the final xgboost model are *nrounds* = 50, *eta* = 0.03 and *max\_depth* = 2. This model achieved an accuracy of 40.07%, sensitivity of 40.03% and specificity of 40.33%.

In attempting to train the XGBoost model, several challenges were encountered regarding the identification of positive classes, even after employing SMOTE to address class imbalance. Despite the synthetic samples generated by SMOTE, the model struggled to accurately identify positive cases, suggesting potential limitations in representing the true distribution of the positive class or the complexity of distinguishing positive from negative cases.

Although the XGBoost final model is not the highest performing of all trials, it offers the best balance between specificity and sensitivity. Using the model with the highest accuracy would result in the failure to identify any positive records, leading to the loss of all distinctive predictive information. While this model does not excel in predictive power, manages to retain some level of sensitivity, ensuring that positive records are not entirely overlooked. Therefore, despite its limitations in predictive accuracy, preserves insights that might otherwise be lost.

***Feature Importance***

Feature importance of the final XGBoost model is assessed based on the contribution of each feature to the overall reduction in the objective function (logloss) during the construction of decision trees. The model calculates feature importance using metrics like *Gain, Cover,* and *Frequency.* *Gain* reflects the improvement in the model's performance achieved by splitting on a specific feature, with higher values indicating greater contribution to predictive power. *Cover* measures a features coverage across observations, indicating how often each feature is used for splitting the data. *Frequency* represents the frequency with which a feature is utilized for data splitting across all trees in the ensemble. These metrics collectively provide insights into the characteristics of the splits made by each feature and their prevalence in the model.

Feature Gain Cover Frequency

1: discharge\_Other 0.82393427 0.72969809 0.66666667

2: preceding\_visits\_binary 0.13854493 0.20028768 0.24666667

3: log\_time\_hosp 0.03752079 0.07001423 0.08666667

*Table 31: Feature importance scores of XGBoost model*

The feature importance scores from the XGBoost model reveal insights into the factors driving the predictive performance of the model. *See Figure 28* to view the first three decision trees in the model. Most prominently, the feature *discharge* *= Other*, indicating whether a patient was discharged to home or elsewhere, stands out with the highest gain score of 0.824. This indicates its substantial contribution to improving the model's accuracy or reducing model error. This feature covers a significant portion of the dataset, approximately 73%, suggesting its prevalence and importance in the prediction process. Moreover, *discharge* *= Other* appears in about two-thirds of the observations, underscoring its frequency within the dataset.

Following closely is the feature *preceding visits binary*, which, while not as influential as *discharge* *= Other*, still makes a noteworthy contribution to the model's performance. With a gain score of 0.139, it demonstrates a moderate impact on predictive accuracy or error reduction. Although it covers a smaller proportion of the dataset compared to *discharge* *= Other* (approximately 20%), *preceding visits binary* remains a relevant factor, appearing in 24.67% of the observations.

In contrast, the feature *log time hosp* emerges with the lowest importance scores among the three. While it still plays a role in the model's predictive capabilities, its impact is relatively minor. With a gain score of 0.038, *log time hosp* contributes modestly to predictive accuracy or error reduction. Additionally, it covers only a small portion of the dataset (approximately 7%) and appears infrequently in the observations, accounting for around 9% of the dataset.

**Random Forest**

The Random Forests (RF) algorithm, pioneered by Breiman (2001)., is an ensemble technique utilized for regression or classification tasks. Its methodology involves constructing numerous individual decision tree models and aggregating their predictions. Input data is distributed to each tree, which then votes for a specific target value based on the majority decision of the individual trees. Subsequently, the forest selects the value with the highest number of votes across all trees.

Using a Random Forest model for classification tasks, the mode is utilized to determine the predicted values by taking the most frequently occurring class label among the predictions made by individual trees. RF leverages Breiman's (2001) bagging approach, coupled with the random selection of predictors for each node, to mitigate overfitting risks. The process of growing each tree entails the following steps:

1. Sampling N records with replacement from the original dataset to obtain the training dataset for the specific tree.
2. Randomly selecting *m* variables from the M input variables, where *m* is less than or equal to M, to split the nodes. This value of *m* remains constant for all nodes.
3. Growing each tree to its maximum depth without pruning.
4. The sole tuning parameters are *m* and *ntree*, where optimal results are achieved when *m* is optimized to reduce correlation between trees.

Through sampling with replacement, approximately two-thirds of the N records are selected, leaving the remaining one-third as the out-of-bag data. This subset serves to obtain an unbiased estimate of the classification error as trees are added. Consequently, RF inherently provides a test dataset, meaning data partitioning and cross validation are not crucial steps for this model.

A diagram of a data flow

Description automatically generated

*Figure 30: The first three decision trees in the XGBoost final model*

A series of Random Forest models were developed for predicting early hospital admission. For each model, the importance of variables was assessed using measures such as LASSO variable importance scores and Gini impurity scores. In Random Forest classification, decision trees make splits based on criteria that optimize the homogeneity or purity of the resulting subsets of data. The split is chosen to minimize the Gini impurity of the resulting nodes, aiming to create subsets with predominantly one class. This process continues recursively until certain stopping criteria are met, such as reaching a maximum tree depth.

​ Random Forest (RF) models were trained and tuned using the unbalanced, balanced and SMOTE-tuned data sets. In the absence of class weights and using the unsampled data, the RF model achieves the highest accuracy (86.25%) but the lowest sensitivity (02.09%). This is likely due to the unbalanced distribution of the outcome variable*.* When utilizing resampled data, where positive records are oversampled to match the same number as negative records, the model exhibits an accuracy of 0.4241, with a sensitivity of 0.68439 and specificity of 0.38456. Notably, there is an enhancement in sensitivity compared to the original unbalanced dataset.

Different combinations of class weights were tested to develop a model that strikes a balance between sensitivity and specificity. When the class weights are set at (1,1), it signifies that negative (majority class) and positive (minority class) records are considered equally important during the training process. As these weight values deviate from (1,1), their significance in the model's decision-making process alters accordingly. Adjusting class weights allows for fine-tuning the model's behavior, with smaller weights indicating less emphasis on correctly classifying positive instances,

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | | | | |
| Data | Class weights | Accuracy | Sensitivity | Specificity | AUC |
| Unbalanced | 1, 1 | 0.8625 | 0.02092 | 0.990336 | 0.50563 |
| Balanced | 1,1 | 0.6179 | 0.46911 | 0.64006 | 0.55458 |
| Balanced | 1,2 | 0.5359 | 0.59837 | 0.52663 | 0.56250 |
| Balanced | 1,5 | 0.4687 | 0.64878 | 0.44183 | 0.54530 |
| SMOTE | 1,2 | 0.6507 | 0.6792 | 0.4593 | 0.56929 |
| SMOTE | 1.3 | 0.5992 | 0.6108 | 0.5211 | 0.56597 |
| SMOTE | 1, 4 | 0.5704 | 0.5577 | .5358 | 0.55564 |
| SMOTE | 1, 8 | 0.5387 | 0.5297 | 0.5992 | 0.56442 |
| SMOTE | 1, 100 | 0.4799 | 0.4572 | 0.6317 | 0.5444 |

*Table 32: Model metrics for contending RF models*

while larger weights prioritize their accurate identification, potentially sacrificing accuracy in negative predictions.

Tuning the RF models with different class weights yields varied results. See *Table 32* for these results. Initially, weighing the minority class with larger values than the majority class lead to an increase in sensitivity, but decrease in specificity as shown by an overall reduced accuracy. Following the criteria for model selection, the optimal model has the highest sensitivity possible while surpassing an accuracy threshold of 60%. To achieve this, lower class weights were explored. Unexpectedly, the most robust model was built using class weights (1,2) meaning the positive class is weighed twice as much as the negative class. Despite integrating SMOTE data into the model, it still faced challenges in correctly identifying positive classes. However, the implementation of class weights helped mitigate some of these issues.

***Variable Importance***

The variable importance scores provide valuable insights into the factors that most strongly influence the predictions made by the random forest model.The *0* column represents the importance score for the variable in predicting the negative class, while the *1* column represents the importance score for the variable in predicting the positive class. The *Mean Decrease Accuracy* indicates how much the accuracy of the model decreases when the variable has its values randomly permuted. Higher values suggest that the variable is more important in maintaining model accuracy. Additionally, *Mean Decrease Gini* is a measure of how each variable contributes to the homogeneity of nodes in decision trees. It assesses the purity of the nodes by evaluating the inequality of class distribution where higher values suggest that the variable is better at separating the classes effectively.

*Table 33: Variable importance scores of final random forest model*

0 1 MeanDecreaseAccuracy MeanDecreaseGini

num\_medications 7.387092 73.13328 61.27568 521.73693

log\_time\_hosp 18.021765 64.61307 49.39098 482.71395

log\_num\_lab 14.358677 62.67251 58.86091 502.30987

preceding\_ts\_binary 8.571578 34.02055 30.53442 60.70403

*log scores* 4.798586 50.18557 41.35072 141.045

age\_30\_to\_60 13.233512 15.41089 32.96455 67.24756

age\_60\_and\_up 14.631217 25.14499 29.86955 85.14580

race\_Caucasian 20.582316 44.73458 38.06474 144.04822

race\_Missing 16.939847 24.20381 28.62645 40.16935

race\_Other 10.428431 22.09249 22.76932 45.74590

*gender*\_Male 27.859950 58.38557 50.06495 179.37980

admission\_type\_id\_Elective 3.994554 30.07724 32.36801 69.7495

admission\_type\_id\_Missing 3.640350 31.78615 30.33014 61.81078

admission\_type\_id\_Urgent 7.792657 43.98863 40.82213 90.893

admission\_source\_id\_Missing 2.563507 21.49168 23.08900 36.12059

admission\_source\_id\_Other 14.213331 48.82704 43.26216 112.45049

admission\_source\_id\_refferal -1.021523 41.58066 38.89702 82.96792

admission\_source\_id\_transfer 18.518468 33.99166 35.59324 67.30364

change\_No 18.665251 43.87854 38.97173 132.49530

diabetesMed\_Yes 28.540589 41.52491 45.76484 133.55133

A1Cresult\_greater\_than\_8 8.201539 19.38108 22.71799 43.717

A1Cresult\_None 15.762310 33.42437 34.21617 83.81213

A1Cresult\_Norm 7.821862 17.56368 19.01102 37.30877

discharge\_Other 49.795256 86.13305 85.71745 255.90142

Because we are interested in the features that predict early readmission, or the positive (1) outcome, the feature importance scores for the *1* column provide valuable insights. For patients who were readmitted, it appears as if *discharge* (86.13)is the most prominent predictor. This is followed by *num medications* (73.13) and *log time hosp (64.61)* and *log num lab (62.67)*. Other notable features are *gender* (58.39), and log scores (50.13).

Assessing the relative importance of these scores compared to the negative class, provides greater insights into their prominence. For instance, the feature *log scores* exhibits a feature importance score of 50.13 for the positive class, contrasting starkly with its score of 4.78 for the negative class. This discrepancy suggests that *log scores* significantly contributes to predicting positive outcomes but holds little relevance in predicting negative outcomes. Similarly, nummedications displays a feature importance score of 73.13 for the positive class, contrasting with its score of 7.39 for the negative class. Surprisingly, log scores and num medications were not emphasized as much in the logistic regression model. However, this disparity could be attributed to differences in how random forests and logistic regression models are able to capture complex interactions within the data.

*A1Cresult = None* demonstrates the highest scores across all HbA1c levels metrics, including *MeanDecreaseAccuracy, MeanDecreaseGini*, and the individual scores for each category. This suggests that the absence of HbA1c test results (i.e., None) significantly influences readmission predictions. In contrast, both *A1Cresult >8* and *A1Cresult = Norm* exhibit lower scores compared to *A1Cresult = None*. However, *A1Cresult > 8* shows slightly higher scores than *A1Cresult = Norm*, indicating that elevated HbA1c levels (>8%) may have a slightly stronger association with readmission risks compared to normal HbA1c levels. Compared to the other predictors, the relatively lower scores for *A1Cresult > 8* and *A1Cresult = Norm* suggest that while HbA1c levels may provide some predictive value for readmission risks, other factors, such as medical history, treatment intensity, and demographic characteristics, may have a more substantial influence on readmission predictions.

Comparing the feature scores between outcomes highlights the significant role of glycemic control and testing in predicting early hospital readmission. For patients who were readmitted, HbA1c testing demonstrates nearly twice the impact on outcomes compared to patients who weren’t readmitted. For instance, *A1Cresult > 8* exhibited a feature importance score of 8.202 for non-readmitted patients, whereas it scored 19.38 for patients who were readmitted. This trend persists across all levels of *A1Cresult*, indicating the crucial predictive value of HbA1c testing in identifying patients at risk of early readmission.

**DISCUSSION**

**Final Models**

Various machine learning algorithms were utilized to model early in-patient hospital readmission, each offering unique perspectives on the model's behavior. Logistic regression excels in providing interpretable results, making it beneficial for understanding the direct influence of individual predictors on readmission likelihood. On the other hand, random forest and XGBoost models are advantageous for capturing complex nonlinear relationships and interactions within the data, thus offering enhanced predictive accuracy. Ultimately, by examining variable importance scores, model coefficients, and odds ratios across these three models, insights into the factors impacting early hospital readmission were gained.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Final Model Metrics** | | | | | |
|  | **Accuracy** | **Sensitivity** | **Specificity** | **Precision** | **AUC** |
| **Logistic Regression** | 0.5911 | 0.63899 | 0.5383 | 0.18860 | 0.611407 |
| **XGBoost** | 40.07 | 0.40034 | 0.40325 | ***0.1984*** | 0.5982 |
| **Random Forest** | 0.6507 | 0.6792 | 0.4593 | 0.14556 | 0.569296 |

#### Table 34: Final Model Metrics

The final logistic regression, random forest and XGBoost model metrics are summarized in *Table 34.* The final logistic regression model achieved an accuracy of approximately 59.11%, a sensitivity of 63.899%, and a specificity of 53.83%. Notably, the AUC-ROC value for this model was 0.6114, indicating fair discriminative ability. The XGBoost model exhibited an accuracy of 40.07%, a sensitivity of 40.034%, and a specificity of 40.325%, with an AUC-ROC value of 0.5982. While its accuracy appears lower compared to other models, it demonstrated a balanced performance across sensitivity and specificity metrics. On the other hand, the random forest model outperformed the others with an accuracy of 65.07%, a sensitivity of 67.92%, and a specificity of 45.93%. Its AUC-ROC value of 0.569296 suggests reasonable discriminative ability. Overall, the random forest model stands out for its higher accuracy and sensitivity, rendering it the most optimal choice.

Throughout this analysis, a persistent tradeoff between model specificity and sensitivity was observed, necessitating multiple evaluations to discern the most robust models. It was noted that optimal models tended to converge around approximately 60% accuracy, 60% sensitivity, and 60% specificity. Despite efforts to enhance accuracy, the resulting models often failed to equally capture information about each readmission outcome, thereby diminishing their utility. This phenomenon underscores the intricate balance required in model development, where the pursuit of the most accurate model must be weighed against the imperative of adequately capturing diverse outcomes.

Various features were engineered to support the predictive ability of the dataset, albeit with varying degrees of success. It was initially expected that comorbidity scores would exert a substantial influence on the model, given their consolidation of valuable patient diagnosis information into a single variable. However, on the contrary, comorbidity scores exhibited minimal impact on the model, characterized by small coefficients, feature importance scores, and odds ratios approaching 1. In contrast, the variable *preceding visits,* representing the number of a patient had over the past 12 months, emerged as a strong predictor, underscoring the advantages of consolidating variable information using domain expertise and insights into variable structure. Notably, preceding visits emerged as one of the most significant indicators in each model.

Data imbalance, a common issue in healthcare datasets, posed significant challenges in this analysis, not only for the outcome variable but also for other predictors. This imbalance can skew model training and compromise the model's ability to accurately identify minority class instances. To mitigate the impact of data imbalance, strategies such as SMOTE, ADASYN, Borderline SMOTE and oversampling positive records were employed. Additionally, class weights were adjusted during model training to place greater emphasis on those of the majority class. Furthermore, careful consideration was given to feature selection and preprocessing techniques to ensure that predictors were informative and representative of both classes. Ultimately, by oversampling the positive records with replacement, the most balanced model and metrics were achieved.

One primary goal of this analysis was to assess the impact of HbA1c levels on early readmission, building upon the findings of Strack et al. (2014). By comparing logistic regression outputs of subsets related to various primary diagnoses, this analysis supported their work by evaluating different profiles among these model summaries. Similar to the findings of Strack et al. (2014), this analysis observed that the likelihood of having an HbA1c test during a patient encounter was higher for patients with a primary diagnosis of diabetes compared to those with respiratory and circulatory diseases.

This analysis extended their work by delving deeper into the profiles of diabetic patients, considering not only those with a primary diagnosis of diabetes but also by diabetes type (controlled, uncontrolled, and unspecified). It was observed that patients with unspecified diabetes had a higher likelihood of early readmission compared to patients with controlled diabetes. This provides insights into the nuanced impact of diabetes management on early hospital readmission rates, highlighting the importance of disease subtypes in predictive modeling and clinical decision-making.

Further expanding upon the research by Strack et al. (2014), this analysis seeks to enhance model performance, a facet not extensively explored in their study. By focusing on improving model performance in terms of predicting positives, a more refined depiction of hospital readmission can be achieved. Several models were meticulously constructed, accompanied by discussions highlighting their respective advantages and limitations. While not all models yielded the anticipated outcomes, experimenting with different models and tuning parameters provided insights into the intricacies of this dataset and the complexities inherent in medical data.

**Implications for Clinical Practice**

***Length of Hospital Stay***

The duration of hospitalization significantly impacts early readmission rates among diabetic patients among all models. Longer hospital stays are associated with an increased risk of readmission, emphasizing the importance of minimizing unnecessary prolonged stays. Healthcare providers should be mindful of the potential risks associated with extended hospitalization and aim to expedite discharge planning wherever possible. This may involve implementing care pathways or protocols designed to facilitate prompt discharge once patients' acute care needs are addressed. Collaborative efforts between healthcare teams and post-acute care providers are essential for ensuring seamless transitions and mitigating the risk of readmission stemming from prolonged hospital stays. By prioritizing efficient discharge processes and optimizing care coordination, healthcare providers can help reduce the likelihood of early readmission and improve overall patient outcomes.

***Laboratory testing***

The frequency of laboratory tests appears to significantly influence early readmission likelihood among all models. Healthcare providers should carefully consider the necessity and timing of laboratory tests, avoiding excessive testing while ensuring essential diagnostic information is obtained. This may involve implementing evidence-based guidelines or protocols for ordering laboratory tests, focusing on tests that provide actionable clinical information relevant to the patient's condition. Additionally, leveraging electronic health record systems to track and monitor laboratory test utilization can help identify opportunities for optimization and reduce unnecessary testing.

***Preceding Visits***

Preceding visits to healthcare facilities emerged as a strong predictor of early readmission across all models. Healthcare providers should pay close attention to patients with a history of frequent healthcare utilization, implementing proactive care management strategies to address underlying health issues and prevent readmissions. This may involve care coordination efforts to ensure continuity of care between different healthcare settings and providers, as well as targeted interventions to address the root causes of frequent healthcare utilization, such as uncontrolled chronic conditions or social determinants of health.

***Demographic Factors***

Patient demographics, such as *age* and *gender*, also play a role in early readmission risk. Healthcare providers should tailor care plans to account for demographic differences, considering factors like age-specific health needs and gender-specific risk factors. This may involve implementing age-appropriate preventive care measures and screenings, as well as addressing gender-specific health concerns or disparities in access to care.

***Admission and Discharge Processes***

Factors related to admission type (admission\_type\_id) and discharge disposition (discharge\_Other) impact early readmission rates, particularly for patients admitted to or discharged from locations other than their home. Healthcare providers should streamline admission and discharge processes, optimizing transitions of care. This may involve implementing standardized admission and discharge protocols, or leveraging technology solutions such as electronic health records and telehealth platforms to help facilitate seamless transitions of care, improve communication between healthcare settings and reduce the risk of readmission due to breakdowns in care coordination.

***Medication and Treatment Compliance***

Medication adherence (diabetesMed\_Yes) and monitoring of HbA1c levels (A1Cresult) are critical for managing diabetes and preventing readmissions. Healthcare providers should emphasize the importance of medication adherence and regular monitoring to patients, providing education and support as needed. This may involve implementing medication management programs, medication reconciliation processes, and patient education initiatives to promote adherence to prescribed treatment regimens.

#### **CONCLUSIONS & LIMITATIONS**

Leveraging data from 54 hospitals in the United States spanning from 1999 to 2008 extracted from the Cerner Health Facts database, various machine learning algorithms were employed to develop classification models predicting readmission probabilities based on a range of clinical and non-clinical variables. This research aimed to investigate the factors influencing early hospital readmission among patients with diagnosed hyperglycemia, with a specific focus on the impact of hemoglobin A1c (HbA1c) levels in combination with other patient demographics, comorbidities, and medication usage.

This analysis was not without its limitations and challenges. Despite extensive efforts to optimize model performance through experimentation with four different machine learning methods and various tuning parameters, none of the models achieved a sensitivity above 60% without compromising accuracy. While the performance metrics of the models consistently hover around the 60% mark, it is important to acknowledge that these metrics, although not exceptionally high, still provide reasonable evidence that the findings could implicate broader issues or serve as a basis for further analysis. However, the inability to surpass the 60% sensitivity threshold suggests that there may be inherent limitations in the dataset or complexities in the relationships between predictors and early hospital readmission among diabetic patients that were not fully captured by the models. Further research exploring alternative modeling techniques or incorporating additional data sources may be warranted to address these limitations.

To address the limitations observed in model performance, it is recommended to employ more extensive tuned logistic regression models that account for variable interactions. By incorporating interactions between variables, these models can capture more nuanced relationships within the data, potentially improving predictive accuracy. Additionally, a Generalized Linear Model (GLM) approach could be considered, where each variable has its own parameter estimate, allowing for a more detailed examination of their individual effects on early hospital readmission among diabetic patients. This approach would enhance the information available in the data and provide insights into the unique contributions of each predictor variable to the outcome of interest. Overall, adopting these advanced modeling techniques could lead to more robust and informative analyses, yielding valuable insights for healthcare providers and decision-makers.

Addressing the challenge of missing data in this analysis requires a multi-faceted approach aimed at minimizing bias and maximizing the reliability of the results. While imputation techniques were employed to handle missing values, variables with a significant number of missing entries posed a challenge to inclusion in the analysis. To mitigate data loss, records with missing values were retained, treating missingness as a distinct category in the analysis. This strategy helps preserve the available information while acknowledging the limitations introduced by missing data. Furthermore, recognizing that the presence of missing data may signal underlying patterns or factors, careful consideration of these instances is essential for accurate interpretation of the results. Moving forward, implementing robust strategies for handling missing data and exploring alternative imputation methods could further enhance the integrity of the analysis and strengthen the validity of the findings.

The non-randomized design of the initial data collection procedure, which may limit the generalizability of the findings also introduces a potential limitation. Strack et al. (2014) extracted data from a comprehensive database to assess HbA1c and other factors associated with diabetes hospital readmissions, but the lack of randomization means that certain patient characteristics or factors influencing readmissions may not be evenly distributed across groups. Despite efforts to mitigate bias through selection criteria and limiting encounters to one per patient, the nonrandomized nature of the study design introduces inherent limitations and may affect the generalizability of the findings.

To address the limitation stemming from the non-randomized design of the initial data collection procedure, researchers might consider conducting further analyses using data from other sources. This can help confirm the robustness of the observed associations and enhance the generalizability of the results. Additionally, efforts can be made to adjust for potential confounding variables through statistical techniques such as propensity score matching or stratification. These methods aim to balance the distribution of covariates between groups, thereby reducing bias and improving the validity of the conclusions drawn from the data. Furthermore, sensitivity analyses can be performed to assess the robustness of the findings under different assumptions or scenarios, providing insights into the stability of the results. Overall, while the non-randomized design poses inherent limitations, careful methodological considerations and analytical approaches can help mitigate these challenges.

In conclusion, this study contributes to advancing research in diabetes care by providing insights into the factors influencing early hospital readmission and offering practical implications for healthcare policies and practices. By informing decision-making processes and guiding interventions, the findings of this research have the potential to improve patient outcomes and drive positive changes in healthcare delivery.

#### **APPENDIX**

|  |  |
| --- | --- |
| **Encounter\_id** | a numerical code assigned to the patient to represent the patient encounter |
| **Patient\_nbr** | a numerical code assigned to the patient to represent patient identification |
| **Race** | the race of the patient |
| ***Gender*** | the *gender* of the patient |
| **Weight** | the weight of the patient (lbs) |
| **Admission\_type\_id** | a numerical code indicating the type of admission (e.g., emergency, urgent, elective) |
| **Discharge\_disposition\_id** | a numerical code representing the patient's discharge disposition (e.g., home, transferred to another hospital) |
| **Admission\_source\_id** | a numerical code indicating the source of the patient's admission (e.g., physician referral, emergency room) |
| **Time\_in\_hospital** | the number of days the patient stayed in the hospital during the current admission. |
| **Payer\_code** | a code representing the type of health insurance or payer for the patient. |
| **Medical\_specialty** | the specialty of the physician providing care to the patient. |
| **Num\_lab\_procedures** | total number of laboratory procedures performed during the current admission. |
| **Num\_procedures** | total number of non-laboratory procedures performed during the current admission. |
| **Num\_medications** | the number of distinct medications prescribed for the patient. |
| **Number\_outpatient** | the number of outpatient visits the patient had in the year preceding the current admission. |
| **Number\_emergency** | the number of emergency department visits the patient had in the year preceding the current admission. |
| **Number\_inpatient** | thenumber of inpatient visits the patient had in the year preceding the current admission. |
| **diag\_1, diag\_2, diag\_3** | primary, secondary, and tertiary diagnosis codes, respectively, indicating the patient's medical conditions. |
| **Number\_diagnoses** | the total number of diagnoses entered for the patient during the current admission. |
| **Max\_glu\_serum** | the result of the maximum glucose serum test during the current admission. |
| **A1Cresult** | the result of the HbA1c test, indicating the average blood sugar level over the past two to three months. |
| **Metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, cetohexamide glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol troglitazone, tolazamide, examide, citoglipton, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone,metformin-pioglitazone** | diabetes management medications prescribed to the patient during the current admission. |
| **Change** | a binary variable indicating whether there was a change in the patient's medications. |
| **diabetesMed** | a binary variable indicating whether the patient was prescribed diabetes medication. |
| **Readmitted** | a categorical variable indicating whether the patient was readmitted to the hospital within 30 days of discharge |

.

#### **REFERENCES**

American Diabetes Association. Complications. (1995-2024). <https://diabetes.org/about-diabetes/complications>

Artetexe, A., Beristain, A., & Graña, M. (2018). Predictive models for hospital readmission risk: A systematic review of methods. Computer Methods and Programs in Biomedicine, 164, 49-64. <https://doi.org/10.1016/j.cmpb.2018.06.006Bliss>, M. (2021). The Discovery of Insulin: Special Centenary Edition. University of Toronto Press.

Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Bruns, D. E., & Knowler, W. C. (2009). Stabilization of glucose in blood samples: why it matters. Clinical chemistry, 55(5), 850–852. <https://doi-org.ccsu.idm.oclc.org/10.1373/clinchem.2009.126037>

Centers for Disease Control and Prevention. (2023). Diabetes Prevention. <https://www.cdc.gov/diabetes/prevention/index.html/>

Centers for Disease Control and Prevention. (2022). Future Surge in Diabetes Could Dramatically Impact People Under 20 in U.S.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357.

Colagiuri, S., Lee, C. M., Wong, T. Y., Balkau, B., Shaw, J. E., Borch-Johnsen, K., & DETECT-2 Collaboration Writing Group (2011). Glycemic thresholds for diabetes-specific retinopathy: implications for diagnostic criteria for diabetes. Diabetes care, 34(1), 145–150. <https://doi-org.ccsu.idm.oclc.org/10.2337/dc10-1206>

Cooke, D. A. (2010). When Your Doctor Says Diabetes: A Guide to Regaining Control Over Your Health and Well-being. Independent Publishers Group.

<https://www.cdc.gov/media/releases/2022/p1229-future-diabetes-surge.html>.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRIS*P-*DM consortium: NCR Systems Engineering Copenhagen (USA and Denmark) DaimlerChrysler AG (Germany), SPSS Inc.\ (USA) and OHRA Verzekeringen en Bank Groep B.V (The Netherlands).

Charlson M, Wells MT, Ullman R, King F, Shmukler C (2014) The Charlson Comorbidity Index Can Be Used Prospectively to Identify Patients Who Will Incur High Future Costs. PLoS ONE 9(12): e112479. <https://doi.org/10.1371/journal.pone.0112479>

Chen, & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Cui, S., Wang, D., Wang, Y., Yu, P. W., & Jin, Y. (2018). An improved support vector machine-based diabetic readmission prediction. Computer Methods and Programs in Biomedicine, 166, 123-135. <https://doi.org/10.1016/j.cmpb.2018.10.012>

Davis, S., Zhang, J., Lee, I., Rezaei, M., Greiner, R., McAlister, F. A., & Padwal, R. (2022). Effective hospital readmission prediction models using machine-learned features. BMC Health Services Research, 22(1), 1415. <https://doi.org/10.1186/s12913-022-08748-y>

Deschepper, M., Eeckloo, K., Vogelaers, D., & Waegeman, W. (2019). A hospital wide predictive model for unplanned readmission using hierarchical ICD data. Computer Methods and Programs in Biomedicine, 173, 177-183. <https://doi.org/10.1016/j.cmpb.2019.02.007>

Demir, E. (2014). A Decision Support Tool for Predicting Patients at Risk of Readmission: A Comparison of Classification Trees, Logistic Regression, Generalized Additive Models, and Multivariate Adaptive Regression Splines. Decision Sciences, 45(5), 849–880. <https://doi-org.ccsu.idm.oclc.org/10.1111/deci.12094>

Dhatariya, K., Corsino, L., Umpierrez, G. E. (2020). Management of Diabetes and Hyperglycemia in Hospitalized Patients. In: Feingold KR, Anawalt B, Blackman MR, et al., editors. *Endotext*. <https://www.ncbi.nlm.nih.gov/books/NBK279093/>

Diabetes Care. (1997). The Expert Committee on the Diagnosis and Classification of Diabetes Mellitus Report of the expert committee on the diagnosis and classification of diabetes mellitus. 20, 1183–1197

Florkowski, C. (2013). HbA1c as a Diagnostic Test for Diabetes Mellitus - Reviewing the Evidence. The Clinical biochemist. Reviews, 34(2), 75–83.

Kelly, M., Longjohn, R., Nottingham, K., The UCI Machine Learning Repository. (2023) <https://archive.ics.uci.edu/about>

Larose, D.T. and Larose, C.D. (2015) Data Mining and Predictive Analytics. Wiley Series on Methods and Applications in Data Mining 2015

Lingjærde, C., Jullum, M. & Sellereite. (2022). shapr: Explaining individual machine learning predictions with Shapley values. <https://cran.r-project.org/web/packages/shapr/vignettes/understanding_shapr.html>

Liu W., Stansbury C., Singh K., Ryan A.M., Sukul D., Mahmoudi E.,Waljee A., Zhu J., Nallamothu, B. K. (2020). Predicting 30-day hospital readmissions using artificial neural networks with medical code embedding. PLoS ONE 15(4): e0221606. <https://doi.org/10.1371/journal.pone.0221606>

Matthews, D., Meston, N., Dyson, P., Shaw, J., King, L., & Pal, A. (2008). Diabetes. OUP Oxford.

Michailidis, P., Dimitriadou, A., Papadimitriou, T., & Gogas, P. (2022). Forecasting Hospital Readmissions with Machine Learning. Healthcare (Basel, Switzerland), 10(6), 981. <https://doi.org/10.3390/healthcare10060981>

O'Connell, J. M., & Manson, S. M. (2019). Understanding the Economic Costs of Diabetes and Prediabetes and What We May Learn About Reducing the Health and Economic Burden of These Conditions. Diabetes Care, 42(9), 1609-1611. <https://doi.org/10.2337/dci19-0017>

Owens-Collins, S. (2023). Beta Cell Regulation, Function, Dysfunction and Eventual Destruction in Type 1 Diabetes. IntechOpen. <https://doi.org/10.5772/intechopen.1001513>

Soundrapandiyan, Rajkumar & Manickam, Adhiyaman & Akhloufi, Moulay & Srinivasa Murthy, Y.V. & Sundaram, Renuka & Thirugnanasambandam, Sivasubramanian. (2023). An Efficient COVID-19 Mortality Risk Prediction Model Using Deep Synthetic Minority Oversampling Technique and Convolution Neural Networks. BioMedInformatics. 3. 339-368. 10.3390/biomedinformatics3020023.

Sugandh, F., Chandio, M., Raveena, F., Kumar, L., Karishma, F., Khuwaja, S., Memon, U. A., Bai, K., Kashif, M., Varrassi, G., Khatri, M., & Kumar, S. (2023). Advances in the Management of Diabetes Mellitus: A Focus on Personalized Medicine. Cureus, 15(8), e43697. <https://doi.org/10.7759/cureus.43697>

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., Clore, J. N. (2014). Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. BioMed Research International, 2014, Article ID 781670. <https://doi.org/10.1155/2014/781670>

The R Project for Statistical Computing. (2015). About The R Project for Statistical Computing. <http://www.r-project.org/about.html>

UCI Machine Learning Repository. (1999-2008). Diabetes 130-US hospitals for years 1999-2008. <https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008>

World Health Organization, & World Health Organization. (2006). International Diabetes Federation. Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation. Geneva: World Health Organization, 1-46.

Yasuhiro Tahara, Kenji Shima; Kinetics of HbA1c, Glycated Albumin, and Fructosamine and Analysis of Their Weight Functions Against Preceding Plasma Glucose Level. Diabetes Care 1 April 1995; 18 (4): 440–447. <https://doi.org/10.2337/diacare.18.4.440>