Predicting Hospital Readmission in Diabetic Patients

Via Paola Dumoran
Master's of Science in Data Analytics
National College of Ireland
Dublin, Ireland
x24151246@student.cirl.ie

***Abstract*— This paper aims at establishing the factors that causes hospital readmission of diabetic patients using logistic regression modelling. After performing exploratory data analysis, several potential key variables were identified including time in hospital, number of medications and HbA1c testing status. The final model included log transformed predictors, polynomial terms and an interaction between A1C levels and the primary diagnosis. The results of the study indicate that longer hospital stays, and absent diabetes management are the major factors that increase the likelihood of readmission. These findings support patient care measures and specific interventions for reducing readmission rates for the future.**

I. Introduction

Hospital readmissions are now considered as one of the biggest problems in the modern healthcare delivery systems world-wide. The high readmission rates represent a system-wide problem, which is often associated with poor discharge planning and follow-up, as well as insufficient treatment. This means that there is something wrong with the quality of care and the process of shifting from hospital to home is a difficult undertaking. This is a significant issue because, not only is it a financial problem for the patient and the health care institution, but it has a significant emotional impact on the patient and their family as it brings stress, uncertainty, and disruption. Furthermore, repeated hospital visits are a burden on the health care systems, strains patients' trust in the health care organizations and has a major impact on the emotional state of the patients.

In order to address this issue, a statistical analysis and logistic regression model was conducted on a dataset developed by Strack et al. (2014) who also performed a similar analysis on diabetic patient records to determine factors that lead to hospital readmissions. They argued that HbA1c testing, which is the test for 'glycated haemoglobin', a blood glucose control parameter and a diabetes indicator, is very low. This resulted in the conclusion that readmission rates were lower with HbA1c testing and that improving diabetes care and surveillance in hospitals decreases readmission rates and the costs of care.

Building on this foundation, my study expands on their work by creating new variables, data preprocessing, removing outliers and enhancing model accuracy through diagnostic evaluation. Since the dependent variable is a binary variable (readmission or not) logistic regression was applied because of its simplicity and interpretability. Hence, the overall purpose is to formulate the best logistic regression model that can be used to predict hospital readmission in diabetic patients and thereby identify the crucial clinical and demographic predictors to enable proactive measures.

II. Data Preprocessing

A. Data Overview

The original dataset contains 101,763 rows and 47 variables before filtering. This includes a mix of categorical and numeric variables like race, gender, age, time in hospital, medications, hospital stay details, and readmission outcomes. The data is generally complete; however, variables such as weight, patient identification, and medications were either excluded, cleaned, or grouped to only include data relevant to the model. After performing some data cleaning, the final row and column count was 100,723 rows and 48 columns.

B. Handling Missing and Inconsistent Data

In order to manage missing or inconsistent variables, several data cleaning steps were applied. Since there are about 23 different drug variables with all the same values of "Up", "Down", "Steady", and "No", they were grouped into a single variable to represent the overall medication status. Although this grouped variable was eventually removed from the final model, the grouping was helpful during the exploratory data analysis. Other variables like weight, patient_nbr, and encounter_id were ignored because of its irrelevance to any predictive values.

Additionally, several categorical values such as gender, race, medical_specialty, admission_source_id, and discharge_disposition were converted into factor variables to ensure they were properly encoded for analysis. To check for missing values, colSums(is.na(df)) was used.

The variable A1CResult, which indicates the result of a patients HbA1c test, included original values such as "None", "Norm", ">7", and ">8". Moreover, a higher value of HbA1c suggests poor blood sugar and a higher risk of complications. To simplify interpretation for analysis, these values were grouped into a new variable A1C_Grouped with values "High HbA1c" (for values ">7" and ">8"), "Normal" for "Norm" and "Not tested" for "None". This simplified interpretation and was used in the final model.

C. Variable Transformation

Not only were certain variables removed, but there were several new variables altered to improve model interpretability. For example, the target variable, readmitted, includes three categorical values: "No" (Not readmitted), "After 30 days", and "Within 30 days". Due to the nature of the project task, the variable was transformed into a binary target which shows whether the patient was readmitted or not.

The hospital dataset uses ICD-9 codes (International Classification of Diseases, 9th Revision) in the diag_1 variable which is the primary diagnosis of a patient. According to the centers for Disease Control and Prevention,

these are unique identifiers for specific illnesses and other health conditions. Therefore, the codes were grouped into a general clinical categories, primary_diagnosis, based on the ranges:

- Between 250 and 259 ➔ Diabetes
- Between 390 and 459 ➔ Circulatory
- Between 460 and 519 ➔ Respiratory
- Between 800 and 999 ➔ Injury/Trauma

D. Outlier Removal

From the boxplots of time_in_hospital and num_medications, it is clear that there are several data points that are significantly different from the rest of the data. In order to reduce any skewness, observations exceeding 1.5 times the IQR above the third quartile were removed and filtered out to ultimately improve the final model quality.

E. Removing Influential Observations

To make a more stable model that better reflects the population, Cook's Distance was used to identify influential observations. Observations with a Cook's Distance greater than 4/n were removed to prevent them from distorting the model results. This ensured that the model's predictions were not excessively dependent on extreme or unusual cases.

III. Exploratory Data Analysis

A. Demographic Analysis

First, the demographic variables were used to initially analyze the population characteristics. From Figure 1, the most common race within the dataset was Caucasian, followed by African American and Hispanic. The remainder remained identified as Asian, Other, or Unknown. In terms of the Gender Distribution in Figure 2, the female patients are slightly higher compared to the male patients. Looking into the age distribution in Figure 3, the majority of patients were within the [60-70) age bracket, indicating a large concentration of middle-aged to older adults among hospital admissions.

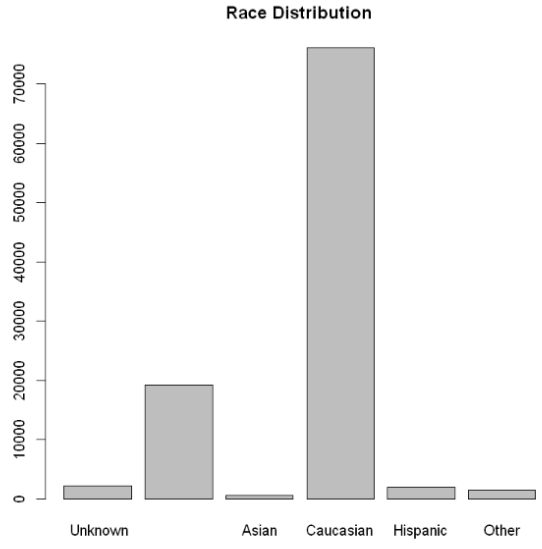

Fig. 1. Distribution of Race within hospital data showing Caucasian has the most common group, followed by African American and Hispanic.

Fig. 2. Distribution of Gender within hospital data showing slightly higher more female patients than male

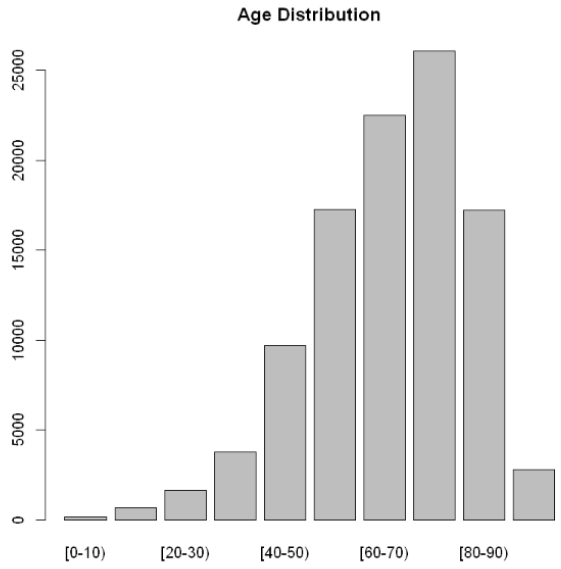

Fig. 3. Distribution of Gender within hospital data showing slightly higher more female patients than male

B. Readmission Distribution

As previously mentioned, the readmitted variable was transformed into readmitted_binary and converted into a factor variable. This step helped to simplify data analysis to help explore the prediction of whether patients admitted or not. Using readmitted_binary, a proportional stacked bar plots were generated to better understand how categorical variables relate to hospital readmissions. These visualisations display the proportion of patients who were admitted vs. not readmitted for each category of age, admission source, HbA1c levels, and medical specialty. Other variables such as gender, race, and diabetes medication status were also explored, but did not show a strong pattern related to readmission.

In Figure 4, the readmission rate by age group varied as middle-aged patients (50-70) had the highest proportion of readmissions, while the youngest (0-20) and oldest (90+) had the lowest. One explanation could be that younger patients tend to have the ability to recover more quickly, reducing the likelihood of readmission. The oldest patients may pass away after being discharged before returning to the hospital, which is not captured as readmission. This also suggests that middle-

aged diabetic patients may undergo more complications or conditions that lead to readmission. This also shows that almost half of the patients who were not tested were readmitted.

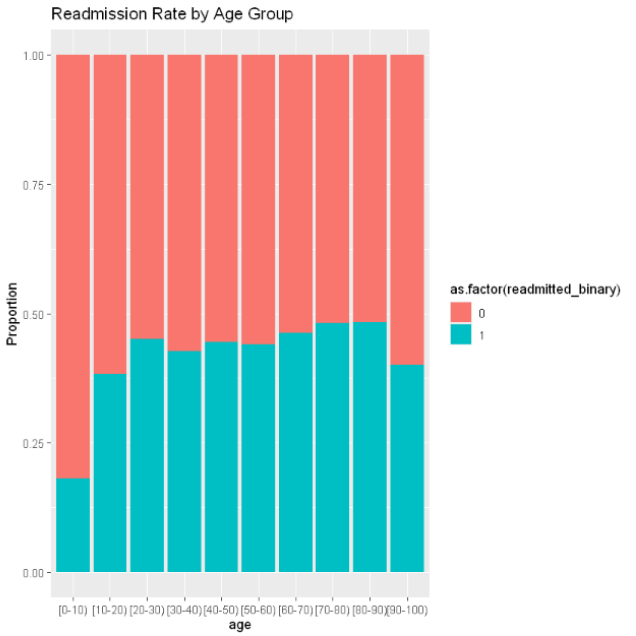

Fig. 4. Proportional readmission rates by age group showing middle-aged patients had the highest proportion of readmissions, while the youngest and oldest age groups showed lower rates.

The article “Impact of HbA1c Measurement on Hospital Readmission Rates” (Strack et al. 2014) provides an assessment of how Hb1Ac testing impacts hospital readmission rates specifically in diabetic patients. From this, we learn that HbA1c testing is rare and only 18.4% of diabetic patients had their HbA1c levels measured. However, when testing is performed, readmission rates were found to be lower. The effect of HbA1c testing on readmission also depended on the primary diagnosis whether it was diabetes, circulatory, respiratory, or any other diagnosis. Moreover, this article highlights that improving diabetes care and testing in hospitals could reduce readmissions. Based on this, this article provides a strong hint that HbA1c measurement is a key predictor of readmission. In Figure 5, the readmission rate by HbA1c level shows that patients with high results had higher rates of readmission, validating that poor diabetic management is a strong predictor of hospital return.

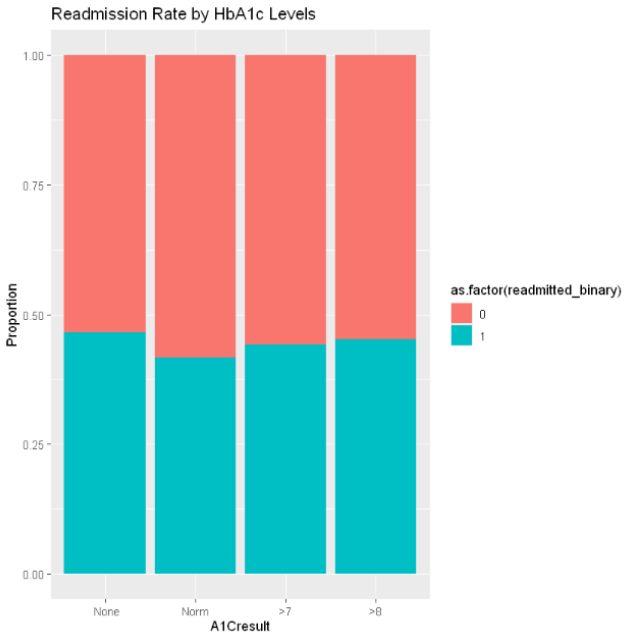

Fig. 5. Proportional readmission rates by HbA1c level showing that patients with high HbA1c levels had higher readmission rates compared to those with normal levels.

In terms of readmission rate by admission source, patients admitted through the emergency room showed higher readmission rates compared to those admitted through other departments shown in Figure 6. This sugggests that emergency admissions which are often asssociated with uncontrolled episodes, which have an increased risk of readmissons emphasizing the need for improved discharged plans and outpatient follow-up. While this was not included in the final model, it provided a better insight into the cases associated with readmissions.

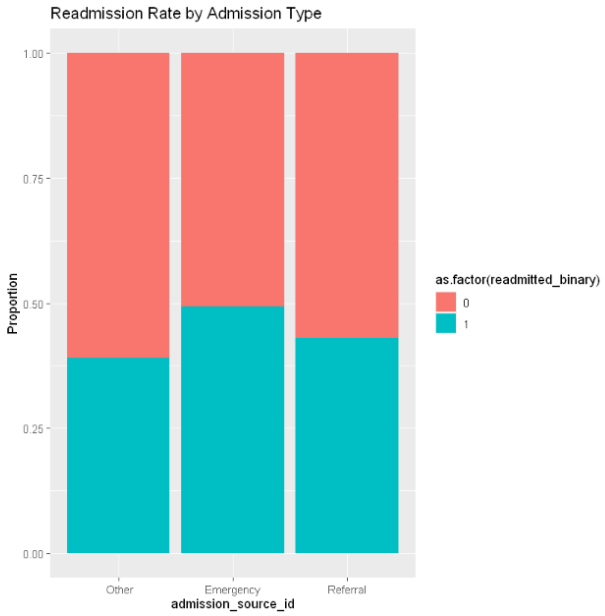

Fig. 6. Proportional readmission rates by admission source showing emergency admissions have the highest likelihood of readmission

Another variable explored was medical speciality which shows the primary physcian or admission service who treated the patient during hospitalization. As shown in Figure 7, patients treated under emergency/trauma and internal medicine had the highest readmission rates. As expected, these departments deal with more complicated and chronic

cases like diabetes complications. While this was also not included in the final model, the visualization provides a better understanding for different environments with higher readmissions.

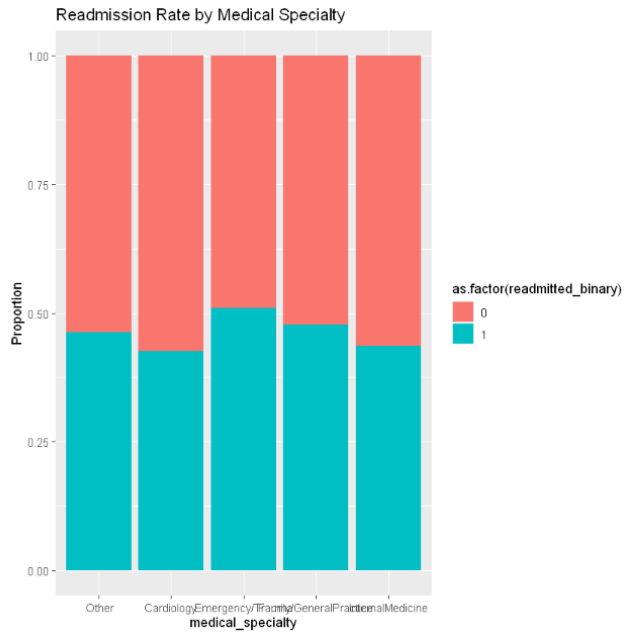

Fig. 7. Proportional readmission rates by medical specialty showing internal medicine and emergency/trauma had the highest rates of readmission

C. Correlation Analysis of Numeric Variables

In order to examine potential relationships, a correlation matrix was generated for all numeric variables. In Figure 8, a correlation heatmap was generated. From this, num_lab_procedures and num_medications showed a moderate positive correlation indicating that patients who went through more lab tests were more likely to receive more medications, which makes logical sense. The variable time_in_hospital was weakly correlated with other numeric variables which shows that the values represent an independent predictor. The overall low-to-moderate correlations confirm that multicollinearity is not a major issue for the numeric predictors.

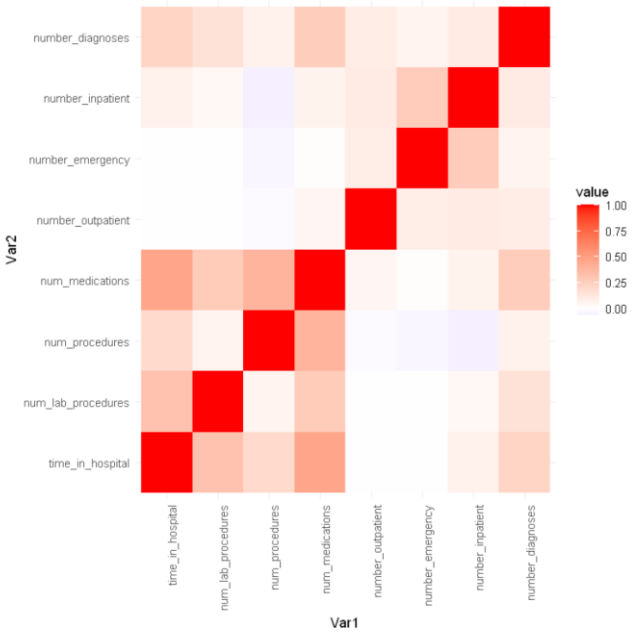

Fig. 8. Correlation heatmap of numeric varibles in the hospital dataset showing low to moderate correlations, indicating low risk of multicollinearity

D. EDA Summary

The exploratory analysis provided several important insights into the patterns and characteristics associated with hospital readmissions among diabetic patients. The demographic analysis revealed that race, gender, and age groups had different distributions, and that middle-aged people were most likely to be readmitted. The key categorical variables of age, HbA1c levels, and admission sources were visually examined as potential predictors that may affect the probability of patients returning to the hospital. The analysis also showed that gender and race did not show strong patterns associated with readmission, but HbA1c test results and admission type showed as stronger predictors.

IV. Logistic Regression Modelling

Following the exploratory data analysis, the next step would be to build a logistic regression model since the correlation matrix displayed low to moderate correlations among numeric variables and the target variable, readmitted_binary, is a binary outcome. This section will focus on building and evaluating a logistic regression model to identify significant predictors leading to an improvement in hospital readmissions.

A. Initial Model Building

To intialize the model building process, the two variables, num_medications and time_in_hospital was used, exluding outliers outside the 1.5*IQR, as previously mentioned in the data cleaning section. These variables were chosen due to the exploratory data analysis showing potential relationship to predicting hospital readmission. Using R, the model was fitted using the glm() function in Figure 9 with readmitted_binary as the target variable.

Model Call					
glm(formula = readmitted_binary ~ time_in_hospital + num_medications, family = binomial, data = df_cleaned)					
Term	**Estimate**	**Std. Error**	**z-value**	**p-value**	**Significance**
Intercept	-0.5344	0.0166	-32.26	< 2e-16	***
time_in_hospital	0.0265	0.0027	9.8	< 2e-16	***
num_medications	0.018	0.0011	17.03	< 2e-16	***
Model Fit Statistics			**Value**		
Null Deviance			134,326 (df = 97,294)		
Residual Deviance			133,680 (df = 97,292)		
AIC	133,686				
Fisher Scoring Iterations	4				
Significance codes: *** p < 0.001, ** p < 0.01, * p < 0.05					

Fig. 9. Summary of logistic regression model with time_in_hospital and num_medications with outliers removed using clean data

B. Varible Selection and Model Refinement

Expanding on the initial model, a categorical variable A1C_Grouped was introduced to enhance the predictive

capability of the model. This variable defines the categories of HbA1c levels, an essential measure of diabetes control, which was identified in the original analysis and explored by previous research (Strack et al., 2014). The variable was converted into a factor to ensure correct treatment in the logistic regression. The ability to include A1C_Grouped in the model enables the consideration of patients' glycemic control as a factor that may affect the risk of readmission. This showed that untreated diabetic patients may lead to more complications and, consequently, more hospital admissions. Hence, the updated model includes one more categorical predictor which are time_in_hospital, num_medications, and A1C_Grouped. Adding one more categorical predictor to the model allows an improved way to determine the relationship between Hb1Ac test results and hospital readmissions. After building this model there was still some room for improvement

This model had two numeric predictors (time_in_hospital, num_medications) and one categorical predictor (A1C_Grouped), with a residual deviance of 140,048 and an AIC of 140,058. Despite all the included variables being statistically significant ($p < 0.001$), the improvement in model fit was fairly modest compared to the null deviance (140,450). In particular, the decrease in deviance suggests some predictive gain, given the large sample size. The AIC is also still fairly high, suggesting that there is potential to improve the performance with either more predictive features or interactions. Also, the small coefficient for num_medications, combined with the marginal effect of the A1C categories, suggests that there is room for improvement using better predictors.

C. *Final Model Summary*

After building multiple models through initializing and adding A1C_Grouped, the interaction between A1C_Grouped and primary_diagnosis was added to test whether the relationship between diabetic testing and readmission depended on the patient's primary diagnosis as previously hinted by Strack et al. (2014). Through this, significant interaction effects were found, showing that patients with specific conditions and difference in HbA1c levels significantly affect the risk of returning to the hospital. Num_medications and time_in_hospital were both log-transformed to reduce skewness, standardize the relationship, and compress extreme values, very similarly to the Simmons logistic regression model example.

As previously mentioned, Cook's distance was used to remove high influence observations to improve the model's stability. Additionally, in order to account for non-linear effects, since most real-world relationships are very rarely linear, squared terms were added for the log_time_in_hospital and log_num_medications. This allowed the model to better capture the relationships between hospital stay and medication load which leads to a more realistic interpretation of the patient population.

Model Call					
glm(formula = readmitted_binary ~ log_time_in_hospital + I(log_time_in_hospital^2) + log_num_medications + I(log_num_medications^2) + A1C_Grouped * primary_diagnosis,					
Term	**Estimate**	**Std. Error**	**z-value**	**p-value**	**Significance**
(Intercept)	-3.813	0.1423	-26.8	< 2e-16	***
log_time_in_hospital	0.608	0.06897	8.82	< 2e-16	***
I(log_time_in_hospital^2)	-0.1311	0.02168	-6.05	1.47E-09	***
log_num_medications	2.166	0.1038	20.87	< 2e-16	***
I(log_num_medications^2)	-0.3768	0.0196	-19.23	< 2e-16	***
A1C_GroupedNorm	-0.5569	0.06849	-8.13	4.28E-16	***
A1C_GroupedNot	0.0966	0.0358	2.7	0.00696	**
primary_diagnosisD	-0.1053	0.05506	-1.91	0.056	.
primary_diagnosisRespiratory	-0.2048	0.06166	-3.32	0.0009	***
A1C_NotTested × Diabetes	0.4284	0.06227	6.88	6.01E-12	***
A1C_NotTested × Respiratory	0.2672	0.06566	4.07	4.72E-05	***
Model Fit Statistics				**Value**	
Null Deviance				130,358	
Residual Deviance				127,476	
AIC				127,518	
Fisher Scoring Iterations				14	
Observations Removed (Cook's D)				1,649	

Fig. 10. Summary of final logistic regression model with polynomial terms and interactions

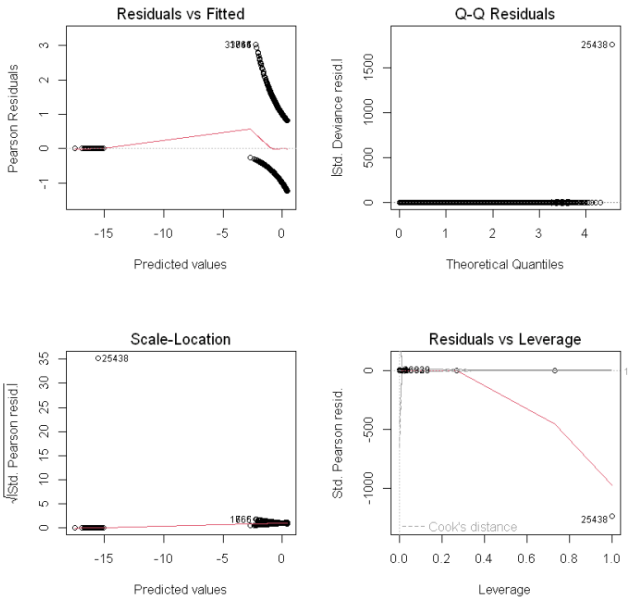

Fig. 11. Diagnostic Plots for the final logistic regression model

D. *Discussion and Interpretation*

According to the model fit statistics of the final logistic regression model, it shows a notable improvement compare to

previous models showing a lower AIC and a larger reduction in deviance. Based on the final model and the predictors, we can see that log_time_in_hospital and log_num_medications had $p < 0.001$ which are significant predictors with a positive coefficient for linear terms. Although this model only shows a 56% accuracy, it is important to note that the accuracy does not represent the full model and had the lowest AIC. This tells us that the odds of readmission increase more with time in the hospital and a greater number of medications.

Analyzing the A1C linked predictors, we realize that patients not tested had increased probability of being readmitted to the hospital, which in turn shows that diabetic control is often poor. Patients who were tested and had normal levels were less likely to be readmitted than patients with high levels. The interaction term A1C_GroupedNot Tested was also statistically significant with Diabetes ($\beta = 0.428$, $p < 0.001$). This indicates that not getting A1C tested is particularly harmful for patients admitted because of problems pertaining to diabetes. Similarly, the interaction with respiratory diagnoses was also significant ($\beta = 0.267$, $p < 0.001$), which indicates some specific diabetes management failures across different diagnoses.

It is important to note that there are several other interaction terms that are not significant, but despite this the model had significant improvement over previous models with lower AIC (127,518). Hence, the model also confirms that longer hospital stays, polypharmacy and poor or absent diabetes care are the major predictors of readmission. The results also highlight the importance of A1C testing, especially in high risk diagnostic categories such as diabetes and respiratory diseases.

V. Conclusion

The purpose of this paper was to determine the factors that lead to hospital readmission of diabetic patients using logistic regression. Age, A1C and admission types were evident in the exploratory data analysis. Through iterative model development, log transformations, polynomial terms, interaction effects, and the removal of influential observations were used to develop the final model that had better fit and interpretability. The results also show that the main drivers of the readmission risk are longer hospital stays, more medications, and poor or absent diabetes care. Moreover, the significance of A1C testing was particularly evident in patients admitted with diabetes and respiratory conditions. These findings stress the importance of more rigorous and specific surveillance and better plans to prevent such occurrences among these populations.

References

[1] StatPearls Publishing, “Hemoglobin A1c,” in *StatPearls [Internet]*. Treasure Island, FL: StatPearls Publishing, Jan. 2024. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK606114I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[2] Health Service Executive (HSE), “HbA1c testing – Managing blood glucose levels,” *HSE.ie*, 2024. [Online]. Available: https://www2.hse.ie/conditions/type-2-diabetes/managing-blood-glucose-levels/hba1c-testing/R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[3] Wikipedia, “List of ICD-9 codes,” *Wikipedia*, Feb. 2024. [Online]. Available: https://en.wikipedia.org/wiki/List_of_ICD-9_codesM. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[4] UCLA: Statistical Consulting Group, “R: Factor variables,” *UCLA Institute for Digital Research and Education*, 2023. [Online]. Available: https://stats.oarc.ucla.edu/r/modules/factor-variables/D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” 2013, arXiv:1312.6114. [Online]. Available: https://arxiv.org/abs/1312.6114

[5] D. Sharma, “Converting Multiple Numeric Variables to Factor at Once in R,” *ListenData*, May 2015. [Online]. Available: https://www.listendata.com/2015/05/converting-multiple-numeric-variables.html

[6] Z. Statology, “How to Use ifelse() in R (With Examples),” *Statology*, 2021. [Online]. Available: https://www.statology.org/ifelse-in-r/

[7] R. W. Nahhas, “Categorical Bar Chart,” in *An Introduction to R*, Bookdown.org, 2020. [Online]. Available: https://bookdown.org/rwnahhas/IntroToR/categorical-bar-chart.html

[8] Z. Statology, “How to Select Only Numeric Columns in R Using dplyr,” *Statology*, 2022. [Online]. Available: https://www.statology.org/dplyr-select-numeric-columns/

[9] R Core Team, “cor: Correlation, covariance matrices,” *R Documentation*, 2024. [Online]. Available: https://www.rdocumentation.org/packages/stats/topics/cor

[10] A. Kassambara, “ggplot2 - Quick correlation matrix heatmap,” *STHDA: Statistical Tools for High-Throughput Data Analysis*, 2021. [Online]. Available: https://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization

[11] Massey University, “Proportional stacked bar graph,” *R Graphics Cookbook*, 2023. [Online]. Available: https://r-resources.massey.ac.nz/rgcookbook/RECIPE-BAR-GRAPH-PROPORTIONAL-STACKED-BAR.html

[12] P. Patel, “Outlier Detection and Removal Using the IQR Method,” *Medium*, Apr. 2023. [Online]. Available: https://medium.com/@pp1222001/outlier-detection-and-removal-using-the-iqr-method-6fab2954315d

[13] J. W. Brown, “icd: Tools for Working with ICD-9 and ICD-10 Codes, and Finding Comorbidities,” *R Documentation*, 2023. [Online]. Available: https://www.rdocumentation.org/packages/icd/versions/4.0.9

[14] R. W. Nahhas, “Binary Logistic Regression: Interaction,” in *Regression Modeling for Public Health*, Bookdown.org, 2020. [Online]. Available: https://www.bookdown.org/rwnahhas/RMPH/blr-interaction.html

[15] C. Thieme, “Logistic Regression with R,” *RPubs*, Mar. 2021. [Online]. Available: https://rpubs.com/christianthieme/76993