A logo for a university

Description automatically generated

**BUSINESS ANALYTICS WITH R**

**Group Project**

**Team 14**

**Course: BUAN 6356.006**

***Hospital Readmissions for Diabetic Patients: A Predictive Modeling Approach***

**Team Members:**

|  |  |
| --- | --- |
| **Darsh Pinal Jogani** | **Shally Preethika Mani** |
| **Thriksha Giriraju** | **Zeeshan Ahmad** |



[Recording of Presentation](https://cometmail-my.sharepoint.com/:v:/g/personal/sxm230262_utdallas_edu/EZ7ce6K2k9JLoDniwPZ6f3EBtvsURVsc6XkK-7N__QMHhA?referrer=Teams.TEAMS-ELECTRON&referrerScenario=MeetingChicletGetLink.view)

Contents

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Topic** | **Page No.** |
| 1. | Executive Summary | 1 |
| 2. | Project Background | 1 |
| 3. | Data Description | 2 |
| 4. | BI Model | 2 |
| 5. | The KDD Process | 3 |
| 6. | Conclusion | 14 |
| 7. | References | 15 |

Executive Summary

Hospital readmission is a growing challenge for the healthcare industry, especially for diabetic patients who are more likely to be readmitted within 30 days (about 4 and a half weeks) of discharge. This project focuses on understanding and predicting early readmissions among diabetic patients using data from 130 US hospitals, covering nearly 70,000 inpatient encounters over ten years (1999–2008). By identifying the key factors influencing readmissions, the aim to help healthcare providers take proactive steps to improve care and reduce costs.

Using advanced analytics and machine learning techniques, such as random forest and multivariate logistic regression, we built predictive models to assess readmission risk. Among the models, Random Forest and Logistic Regression performed the best, achieving an impressive accuracy of 90.86% and 90.72%. These results demonstrate the model’s ability to reliably identify patients at risk of early readmission.

Our analysis identified critical factors that contribute to readmissions, including the type and source of admission, the number of diagnoses and procedures, and the nature of the primary diagnosis. These findings provide valuable insights for hospitals and healthcare professionals to address the root causes of readmissions. By using these insights from this project, hospitals can implement targeted interventions for high-risk patients, such as adjusting treatment plans, conducting more thorough diagnoses, or ensuring follow-up care. This can lead to better health outcomes for patients, lower readmission rates, and reduced medical costs.

In summary, this study highlights how predictive analytics can play a vital role in improving healthcare for diabetic patients. The results not only provide actionable insights but also pave the way for developing effective strategies to enhance patient care and optimize resources in the healthcare system.

Project Background

Diabetes is a leading global health challenge, with its prevalence doubling over the past two decades. Managing this chronic condition is complex, often requiring ongoing care and monitoring. Hospital readmissions, especially within 30 days of discharge, are a critical issue for diabetic patients, signaling gaps in care and placing a significant burden on healthcare systems.

Diabetic patients face a higher risk of readmission due to the multifaceted nature of the disease and its associated complications. These readmissions drive up healthcare costs and negatively impact patients’ quality of life, making it crucial to identify and address contributing factors.

This project is motivated by the need to reduce readmissions for diabetic patients, improving their outcomes and lowering medical costs. Using data from 130 US hospitals, covering nearly 70,000 inpatient encounters, the study aims to uncover key predictors of early readmissions and build predictive models to assist healthcare providers. By leveraging insights from factors such as admission type, number of diagnoses, and procedures, this project demonstrates the power of data-driven approaches in enhancing diabetes care and addressing broader challenges in healthcare.

Data Description

This project utilizes the **Diabetes 130-US hospitals dataset**, which contains information on diabetic patient encounters collected from 130 hospitals and healthcare systems between 1999 and 2008. The dataset includes over 80,000 records with more than 50 features, covering patient demographics, clinical details, and hospital-related outcomes. It serves as a valuable resource for studying factors that contribute to hospital readmissions among diabetic patients. Below is the link to the original dataset:

<https://doi.org/10.24432/C5230J.>

**Key Dataset Features:**

* **Demographics**: Includes attributes such as patient age, race, and gender.
* **Admission Details**: Covers variables such as type of admission (e.g., emergency or elective), length of hospital stays, and discharge destination.
* **Medical Information**: Includes test results (e.g., HbA1c), primary and secondary diagnoses, number of lab tests conducted, and types of diabetes-related medications prescribed.
* **Visit History**: Tracks outpatient, inpatient, and emergency visits within the year prior to hospitalization.

**Target Variable:**

The primary label used for prediction in this study is **readmission status**, categorized into three groups:

* **No readmission**: The patient did not return to the hospital.
* **Readmission within 30 days**: Indicates potential inadequacies in treatment.
* **Readmission after 30 days**: May reflect the patient’s underlying condition rather than immediate treatment issues.

The focus of this analysis is predicting **readmissions within 30 days**, as this category has significant implications for improving treatment quality and patient outcomes.

BI model

We used two models to predict whether a patient would be readmitted: Random Forest and Logistic Regression.

* **Random Forest** combines multiple decision trees to improve accuracy and reduce the risk of overfitting.
* **Logistic Regression** provides a statistical approach, focusing on the relationship between features and the probability of readmission.

After training the models, we evaluated their performance to identify the most effective method for this classification problem.

The KDD Process

A diagram of data mining

Description automatically generated

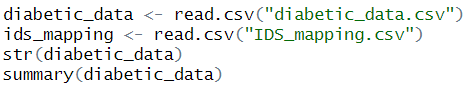
The following steps which are part of the *Knowledge Discovery in Databases* are being followed in this project:

**Step 1: Data Selection**

The data selection process focused on identifying and importing the datasets required for the analysis. The primary dataset, **diabetic\_data**, was chosen as it contains detailed information relevant to the case study objectives. A supplementary dataset, **IDS\_mapping**, was also considered for potential use in future stages but was not utilized at this point.

The **read.csv ()** function is used to import datasets into R, converting CSV files into data frames for easier analysis. To understand the structure and organization of the imported data, the **str ()** function is applied, which provides details such as column names, data types, and a glimpse of the values within the dataset. Additionally, the **summary ()** function is utilized to generate key summary statistics, including averages, medians, ranges, and counts, offering a quick overview of the dataset's numerical and categorical attributes. This step ensured that the target dataset was ready for the next phases of the analysis, including cleaning and transformation.

The code written for this step is as follows:



A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

**Step 2: Data Preprocessing**

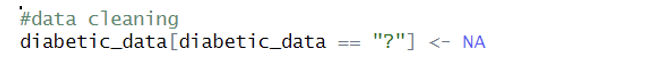
Data preprocessing is a foundational step in preparing raw datasets for analysis and machine learning. The primary goal is to clean, transform, and format data into a usable structure, ensuring high data quality and compatibility with analytical tools and algorithms. The steps involved in preprocessing address issues such as missing data, inconsistent formatting, and irrelevant or redundant features.

#### **Handling Missing Values**

Missing data is a common issue in real-world datasets, and it can significantly impact the accuracy of analysis and models.

In this process, we are:

1. **Replacing Placeholder Values:** Placeholder values such as "?" are present in this dataset and they are replaced with NA to standardize missing data representation and facilitate further handling.



1. **Mapping Invalid IDs to NA:** Certain values in categorical columns (admission\_type\_id, discharge\_disposition\_id, and admission\_source\_id) are recognized as placeholders or invalid entries and are converted to NA. This is found by using the IDS Mapping. We will be able to find the IDS mapping CSV file under the dataset files section in this link. <https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008>

#### A close up of text Description automatically generated

#### **Data Type Conversion**

Converting columns to appropriate data types ensures that they are interpreted correctly by analytical tools and models:

* **Categorical Columns:** Variables such as race, gender, age, etc and categorical IDs are explicitly converted to factor data types.

A screenshot of a computer code

Description automatically generated

This step ensures that categorical variables are correctly treated in statistical analysis and machine learning workflows.

#### **Data Filtering**

Filtering reduces noise and irrelevant entries, refining the dataset:

**Selective Retention:** Rows where discharge\_disposition\_id corresponds to hospice or death-related cases are retained using filtering conditions.

A close-up of a data

Description automatically generated

This ensures that only relevant data points remain for analysis, improving model focus and computational efficiency.

#### **Feature Pruning**

Feature pruning focuses on identifying and retaining only the most relevant variables while discarding redundant or irrelevant ones:

* It reduces dimensionality, simplifies the dataset, and minimizes overfitting risks.
* Pruning can involve techniques such as correlation analysis or removing low-variance features.

In this case, irrelevant or placeholder IDs are handled during missing value replacement and filtering.

A computer code with text

Description automatically generated

1. **Exploratory Data Analysis (EDA)**

Exploratory Data Analysis (EDA) is a vital step in this process and is performed to understand the dataset's structure, distributions, and relationships. This step is performed to gain initial insights into a dataset, understand its structure, and identify patterns, trends, and anomalies. It helps detect potential issues in the data, and guide further statistical analysis or machine learning model development.

Key results from the EDA conducted on the diabetic\_data dataset is detailed below:

* 1. **Missing Value Analysis**
* The **proportion of missing values** is visualized for all features using a bar chart.
* Key observations are as follows:
* Features with a high proportion of missing values may require imputation or removal.
* The chart helps prioritize columns for cleaning based on missingness rates.

A graph with text and numbers

Description automatically generated with medium confidence

The above image shows the proportion of missing values for every feature.

**2. Distribution of Categorical Variables**

Histograms are used to visualize the distribution of categorical variables such as:

**Demographics:**

* **Race:** Most patients are Caucasian, with other racial categories forming smaller proportions.
* **Gender:** A balanced distribution is observed between male and female patients.
* **Age Groups:** A majority of patients belong to the older age brackets.

**Medical Data:**

* **A1C Results:** Captures the frequency of patients with abnormal A1C levels.
* **Admission Types:** Provides insights into the reasons for hospitalization.
* **Discharge Status:** Shows the proportion of patients discharged to different care facilities.

These visualizations help identify imbalances or potential biases in the data, which can inform preprocessing and analysis.

A computer screen shot of a program code

Description automatically generated

A screenshot of a graph

Description automatically generated

**3. Distribution of Numerical Variables**

The numerical features, such as time\_in\_hospital, num\_lab\_procedures, and glucose measurements, are plotted to observe their distributions:

* **Histogram Insights:**
  + Numerical variables generally follow skewed or non-normal distributions.
  + Features such as glucose levels may require transformations (e.g., log-scaling) for modeling purposes.

A computer screen shot of a program

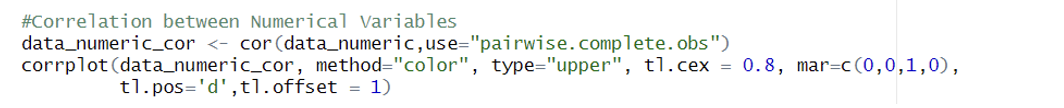
Description automatically generated

A graph of a diagram

Description automatically generated with medium confidence

**4. Correlation Analysis**

* A correlation matrix is used to explore relationships between numerical variables, visualized with a heatmap:
* **Key Correlations:**
* time\_in\_hospital shows the strongest positive correlation with num\_lab\_procedures.
* Most other variables exhibit weak or negligible correlations, indicating relatively independent relationships between different hospital visit measures.



A blue squares with red text

Description automatically generated

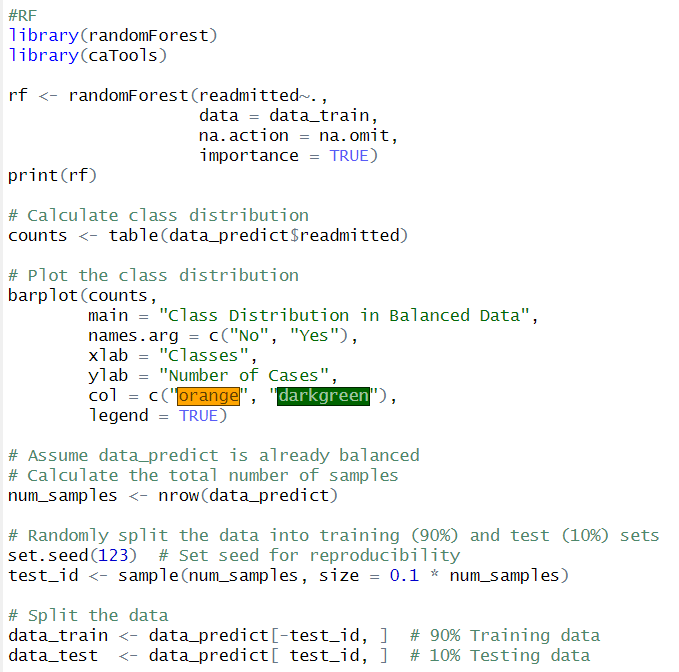
**Step 3: Data Mining**

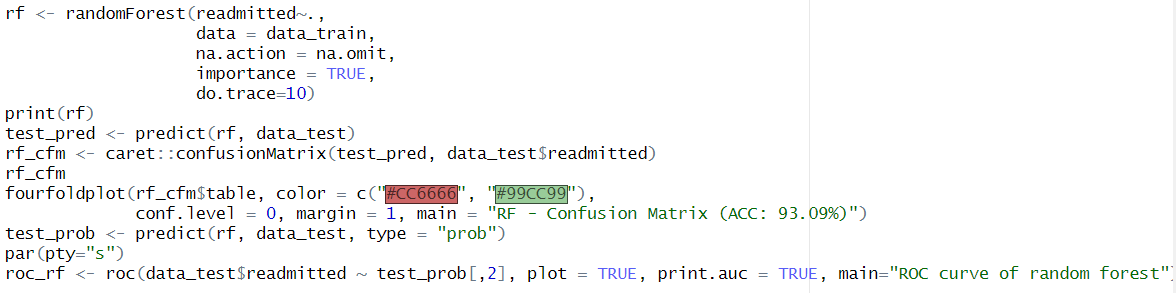
Data mining is a powerful process that extracts valuable insights from large datasets using statistical analysis and machine learning techniques. It helps organizations uncover hidden patterns, relationships, and trends to support informed decision-making and strategic planning. Data mining is widely adopted across industries such as retail, healthcare, finance, and manufacturing, where it aids in solving complex business problems and improving operational efficiency.

In this project, two data mining techniques were used – Random Forest and Logistic Regression, to predict the probability.

Random Forest is a powerful ensemble learning algorithm that combines multiple decision trees to make predictions, improving accuracy and reducing overfitting. It excels in handling complex datasets with high dimensionality, automatically selecting important features, and providing robust performance across various classification and regression tasks.

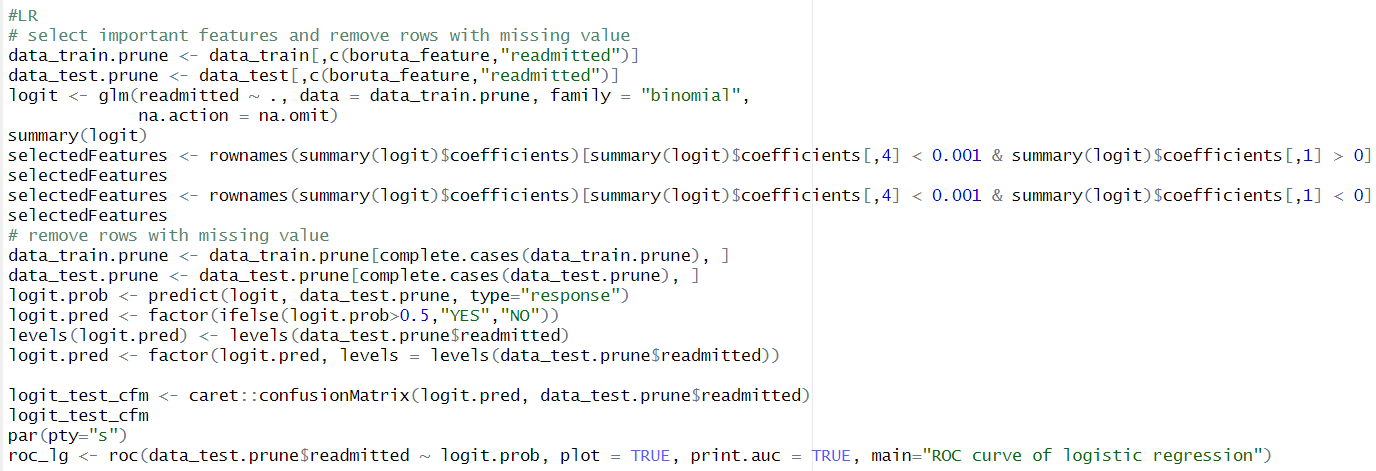
Code for Random Forest -





The next model used was Logistic Regression.

Logistic Regression is a statistical method used for predicting binary outcomes by estimating the probability of an instance belonging to a particular class. It's widely used in various fields for classification tasks, offering interpretable results and serving as a foundation for understanding more complex machine learning algorithms.



**Step 4: Interpretation/Evaluation**

Results and Analysis:

These are some of the results derived:

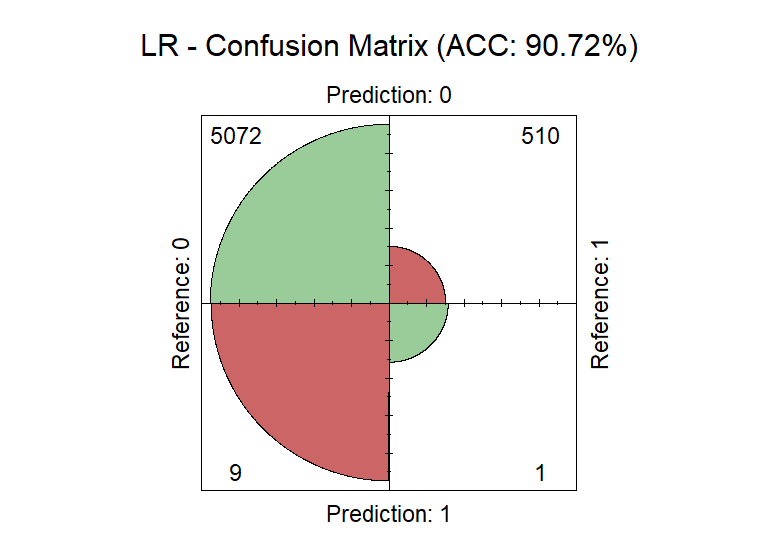
* **Accuracy**: Both models performed similarly in terms of overall accuracy, with Random Forest slightly outperforming Logistic Regression (90.86% vs 90.72%).
* **Sensitivity**: Random Forest achieved perfect sensitivity (1.00), meaning it correctly identified all positive cases (patients who were readmitted). Logistic Regression was also highly sensitive (0.9982).
* **AUC (Area Under the Curve)**: Logistic Regression had a slightly higher AUC (0.674) compared to Random Forest (0.659), indicating a marginally better overall discriminative ability.
* **ROC Curves**: The ROC curves for both models were plotted, visually confirming the similar performance with a slight edge for Logistic Regression.

A green and red diagram

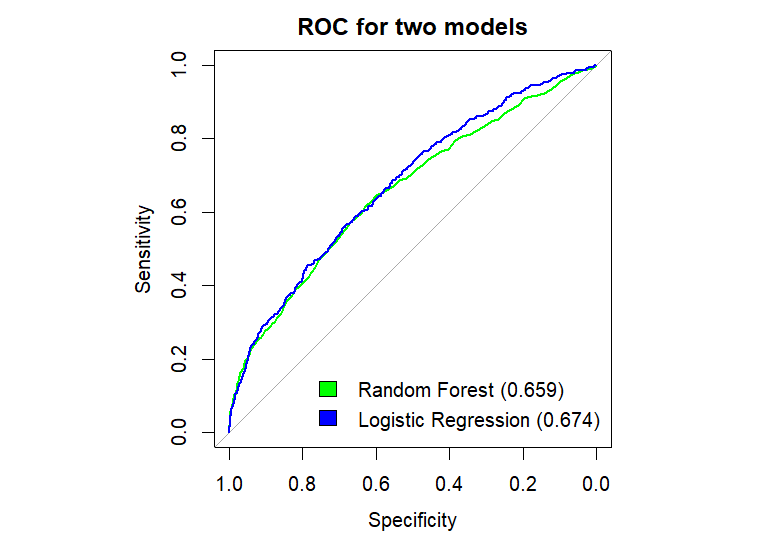
Description automatically generated

The above image shows the Confusion Matrix for the Random Forest. The accuracy achieved was 90.86%.

The below image shows the Confusion Matrix for the Logistic Regression. The accuracy achieved was 90.72%.



The below image shows the comparison of the ROC curves for both the models.



Conclusion

**Predictive Model Performance**

Both Random Forest and Logistic Regression models demonstrated high accuracy (90.86% and 90.72% respectively) in predicting hospital readmissions within 30 days.

The models showed excellent sensitivity (1.00 for Random Forest and 0.9982 for Logistic Regression), indicating they are highly effective at identifying patients at risk of readmission.

**Key Factors Influencing Readmission**

The feature selection method (Boruta) identified several important predictors of readmission. These are:

* Patient demographics (e.g., age, race)
* Admission details (e.g., admission type, discharge disposition)
* Diagnosis groups
* Laboratory test results (e.g., A1C levels)
* Medication changes

**Business Implications**

1. *Resource Allocation*: The high predictive accuracy of the models can help hospitals better allocate resources by identifying patients most likely to be readmitted.
2. *Targeted Interventions*: Healthcare providers can use the model predictions to implement personalized care plans for high-risk patients, potentially reducing readmission rates.
3. *Cost Reduction*: By preventing unnecessary readmissions, hospitals can significantly reduce healthcare costs and improve overall efficiency.
4. *Quality of Care*: Focusing on factors that contribute to readmission can lead to improvements in the overall quality of care provided to diabetic patients.
5. *Patient Segmentation*: The clustering analysis reveals distinct patient groups, which can be used to tailor treatment approaches and follow-up care strategies.

**Recommendations**

1. Implement a predictive model in the hospital's electronic health record system to flag high-risk patients for special attention.
2. Develop targeted intervention programs focusing on the key factors identified by the models.
3. Enhance discharge planning and post-discharge follow-up for patients identified as high-risk for readmission.
4. Conduct further analysis on the identified patient clusters to understand their unique characteristics and needs.
5. Regularly update and refine the predictive models with new data to maintain their accuracy and relevance.
6. By leveraging these insights, healthcare providers can work towards reducing readmission rates, improving patient outcomes, and optimizing resource utilization in the management of diabetic patients.

References

1. Strack, Beata, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. 2014. “Impact of Hba1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records.” *BioMed Research International* 2014.
2. Data Mining for Business Analytics: Concepts, Techniques,and Applications in R, by Galit Shmueli, Peter Bruce, Inbal Yahav, NitinPatel, and Kenneth Lichtendahl. Wiley, ISBN-10: 1118879368, ISBN-13:978-1118879368
3. Clore, J., Cios, K., DeShazo, J., & Strack, B. (2014). Diabetes 130-US Hospitals for Years 1999-2008 [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5230J>.