# **Analytics Startup Plan**

**Synopsis: *This document provides a high-level walkthrough of the activities required to guide completion of the analysis.***

|  |  |
| --- | --- |
| **Project** | Hospital Readmission Analysis for Diabetic Patients |
| **Requestor** | Mohamed Thalha Ahamed Ali |
| **Date of Request** | Jul 14, 2025 |
| **Target Quarter for Delivery** | Q3, 2025 |
| **Epic Link(s)** | N/A – academic capstone |
| **Business Impact** | This project will help the hospital identify why some diabetic patients are readmitted within 30 days and enable targeted interventions before discharge. Reducing avoidable readmissions will lower costs, ease pressure on staff and beds, improve patient outcomes, and strengthen the hospital’s performance scores and reputation. |

## **1.0 Business Opportunity Brief**

|  |  |
| --- | --- |
|  | Clearly articulated business statement of the Ask, opportunity, or problem you are trying to solve for. An important step is to understand the nature of the business, system or process and the desired problems to be addressed. This will be communicated back to All stakeholders for alignment. |

**Problem Statement:**

High 30-day hospital readmission rates among diabetic patients increase healthcare costs, strain limited hospital resources and lower the quality of patient care. These unplanned returns also expose hospitals to financial penalties under value-based care programs. Without an effective way to identify and intervene for high-risk patients before discharge, the cycle of preventable readmissions will continue.

**Opportunity:**

Hospital readmissions within 30 days are a costly, measurable indicator of gaps in post-discharge care and patient management. Leveraging historical hospital encounter data for diabetic patients offers a chance to uncover patterns and key drivers of readmission risk. By applying advanced analytics and machine learning, the hospital can target high-risk individuals with proactive follow-up care, optimize the use of staff and bed capacity, and improve patient outcomes while reducing financial losses.

**The specific ask:**

Design and deliver an end-to-end analytics pipeline—covering exploratory data analysis, feature engineering, class balancing, and predictive modeling—that identifies the most influential factors driving <30-day readmissions among diabetic patients. The final output should include an interpretable, high-recall model that can be integrated into the hospital’s workflow to flag at-risk patients before discharge, along with actionable insights for clinical teams to guide targeted interventions.

## **1.1 Supporting Insights**

|  |  |
| --- | --- |
|  | Define any supporting insights, trends and research findings. Where relevant, list key competitors in the market. What are their key messages, products & services? What is their share of market, nationally and regionally? |

Diabetic patients face a significantly higher risk of early hospital readmissions compared to the general population due to the chronic nature of the disease, frequent comorbidities, and the complexity of post-discharge management. Studies indicate that nearly one in five diabetic patients is readmitted within 30 days, often for preventable complications.  
Government and insurance programs worldwide are increasingly linking hospital funding to readmission rates, penalizing institutions that exceed benchmarks. As a result, reducing early readmissions has become a top priority for hospitals seeking to maintain financial stability and improve quality scores.  
  
**Link:** <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7418689/>

## **1.2 Project Gains**

|  |  |
| --- | --- |
|  | *Describe any revenue gains, quality improvements, cost and time savings (as applicable). What will you do differently and why would our customers care. What are the implications if we do nothing? This section is particularly key for prioritization against company goals and KPI’s.* |

**Gains:**

* **Better Patient Outcomes:** Early identification of at-risk patients enables targeted interventions such as follow-up calls, specialist referrals, or medication adjustments.
* **Cost Savings:** Preventing avoidable readmissions lowers treatment costs and reduces the financial penalties associated with value-based care programs.
* **Improved Hospital Efficiency:** Fewer repeat admissions free up beds and staff for new cases, improving operational capacity.
* **Reputation and Quality Scores:** Demonstrating effective readmission reduction supports better public quality ratings and trust from the community.

**Risk of Doing Nothing:**

If no action is taken, the hospital will likely continue to face high readmission rates, elevated costs, increased strain on resources, and potential financial penalties missing the opportunity to improve both patient outcomes and operational performance.

## *Note: Completion of the following sections is possible only after a careful assessment and triage of the Ask. This is required to determine scope, resource, time, priority and data availability.*

## **2.0 Analytics Objective**

|  |  |
| --- | --- |
|  | List the key questions, assumptions and define the hypotheses. Often the deliverable may not just be an analysis output, however a recommended operating model or blueprint for a pilot etc.  Note: Asking the right questions and truly understanding the problem will lead to the right data, right mathematics, and right techniques to be employed. |

The objective is to develop a predictive model capable of identifying diabetic patients at high risk of being readmitted within 30 days of discharge. This will involve exploring historical hospital encounter data to uncover the most influential factors, engineering clinically meaningful features, and applying machine learning models optimized for recall and interpretability.  
The final deliverable will include:

* A ranked list of key predictors of early readmission.
* A validated model suitable for integration into hospital workflows.
* Actionable recommendations for targeted post-discharge interventions.

### **Key Questions:**

* What demographic, clinical, and administrative variables appear most relevant to early readmissions?
* How do hospital utilization patterns (prior inpatient, outpatient, and emergency visits) relate to future readmissions?
* Are there notable differences in readmission risk across diagnosis groups or discharge dispositions?
* How do missing values and grouped categories affect the overall analysis?
* What is the distribution and balance of readmission cases, and how should it be handled in modelling?

## **2.1 Other related questions and Assumptions:**

|  |  |
| --- | --- |
|  | *List any assumptions that may affect the analysis* |

* Does length of stay correlate with readmission likelihood?
* Are there specific combinations of diagnoses and medication changes that indicate higher risk?
* How does the number of prior hospital visits influence a patient’s probability of being readmitted?
* Are certain admission sources linked to increased readmission rates?

**Assumptions:**

* The dataset is representative of real-world patient encounters and is sufficiently complete for analysis.
* Grouping ICD9 codes into broader diagnosis categories preserves predictive power while improving interpretability.
* Each hospital encounter is treated independently, even for patients with multiple visits.

## **2.2 Success measures/metrics**

|  |  |
| --- | --- |
|  | *What does success look like? Define the key performance indicators (success definition/indicators, drivers and key metrics) against which the objectives will be analyzed. These should be drawn from the interlock meeting with key stakeholders and will inform the approach and methodology for the analysis.* |
|  | **Key Indicators of Success:**   * Clearly identifying a set of factors most associated with 30-day readmission risk through EDA and statistical testing. * Building models that achieve acceptable performance in recall (primary) and ROC-AUC (secondary) while remaining interpretable. * Producing findings that are actionable for hospital teams and can guide follow-up strategies.   **Key Metrics to Track:**  Success was defined using:  - Recall: Critical to minimize false negatives (high-risk patients missed).  - ROC-AUC: Ensures model ranking quality. |
|  |  |

## **2.3 Methodology and Approach**

|  |  |
| --- | --- |
|  | *Now that you have a good understanding of the Ask and deliverable, detail the recommended approach/methodology.* |

**Type of Analysis:**

The analysis will be conducted in structured stages to ensure that exploratory findings, data preparation, and modeling are all aligned with the project objective of identifying factors related to 30-day readmission in diabetic patients.

**Methodology:**

**1. Data Understanding & Initial Checks**

* Reviewed dataset structure, data types, and completeness using descriptive statistics and frequency checks.
* Quantified distribution of the target variable (<30-day vs No readmission) to assess class imbalance before modelling.

**2. Target Variable Definition**

* Created three possible binary target variables to test different problem framings:
  1. No vs. <30-day
  2. No vs. >30-day
  3. <30-day vs. >30-day
* Compared performance across all three definitions and selected Option 2 (No vs. <30-day) to focus on immediate readmissions and exclude long-gap noise.

**3. Data Cleaning & Preparation**

A screenshot of a computer

AI-generated content may be incorrect.

* Dropped columns with extremely high missingness (Weight, ~97% missing).
* Removed constant-value drug columns (examide, acetohexamide, glimepiride-pioglitazone) as they offered no variance.

A blue square with black text

AI-generated content may be incorrect.A blue rectangular object with white text

AI-generated content may be incorrect.

A graph with a blue square

AI-generated content may be incorrect.

* Imputed missing categorical values:

Race → mode imputation.

Payer\_code, Medical\_specialty, Diagnosis fields → “Missing” or “Unknown” category.

* Created missing indicators for Payer\_code, Medical\_specialty, Diagnosis fields, and Weight.
* Mapped Diag\_1, Diag\_2, Diag\_3 ICD9 codes to grouped diagnosis categories using a standard ICD reference file.
* Created a numeric **age midpoint** variable from the age range categories.

**4. Grouping High Cardinality Categorical Variables**

* **Admission Type ID** → Emergency/Urgent, Elective, Other.
* **Discharge Disposition ID** → Home, Transferred, Deceased, Other.
* **Admission Source ID** → Referral, Emergency, Transfer, Other.
* **Diagnosis Groups** → Circulatory, Diabetes, Respiratory, Digestive, Injury, Other.
* **Medical Specialty** → Internal Medicine, Cardiology, Surgery, Orthopedics, Other.

**Skewness Check for Numeric Variables:**

Checked skewness statistics for all numeric features to ensure they follow a near-normal distribution, which improves model performance.

**Highly Skewed Variables:**

number\_emergency, number\_inpatient, and number\_outpatient showed high positive skewness (long tails with extreme values).

Applied Log2 transformation to reduce skewness. This transformation successfully reduced skewness but did not affect the outlier count for number\_emergency and number\_outpatient.

A screenshot of a computer

AI-generated content may be incorrect.

**Final Approach:**

Skipped Log2 transformation for number\_emergency and number\_outpatient because outliers persisted.

Capped outliers for all other skewed numeric variables (using the IQR cap method) to prevent extreme values from influencing the model.

A computer screen shot of a computer code

AI-generated content may be incorrect.

**6. Interaction Features:** Created clinically meaningful interaction terms:

- Stay × Medications (proxy for treatment intensity)

- Inpatient × Emergency visits (acute instability indicator)

- Age × Diagnoses (chronic complexity measure)

- Procedures × Stay (care intensity).

**7. Encoding & Split:**

* One-hot encoded all categorical variables.
* Applied a 70/30 stratified train-test split to maintain class proportions.

**8.** **Class Balancing:**

* Compared multiple balancing techniques on the **training set only** to avoid data leakage:
  + Random Oversampling (ROS)
  + Random Undersampling (RUS)
  + SMOTE
  + Tomek Links (specifically for Logistic Regression with SelectKBest variable selection).

**9. Modeling:**

* Tested multiple algorithms: Decision Tree, Random Forest, Logistic Regression, XGBoost, AdaBoost, Neural Network.
* Performed hyperparameter tuning for selected models (e.g., tree depth for Decision Tree, n\_estimators and max\_depth for Random Forest, architecture and dropout rate for Neural Network).
* For Logistic Regression, applied **Tomek Links + SelectKBest** to address imbalance and select significant predictors.

**10. Evaluation:**

Evaluated models using **Recall** (primary metric) to minimize false negatives and **ROC-AUC** (secondary metric) for ranking quality.

## **3.0 Population, Variable Selection, considerations**

|  |  |
| --- | --- |
|  | Capture learning about the data available today location, structure, and reliability; this would include data in operational systems including dealer sourced, data warehouse and any CRM or email marketing systems available today. |

**Audience / Population Selection:**  
The analysis focuses on hospital encounters involving patients diagnosed with diabetes, as recorded in the UCI Machine Learning Repository dataset (1999–2008, 130 hospitals in the United States). Each observation represents a single hospital encounter; some patients appear multiple times due to repeat visits.

**Observation Window:**

* Target: whether the patient was readmitted within 30 days of discharge.
* Encounters with “>30 days” readmission were excluded from the final modeling target.

**Inclusions:**

* Records with a valid readmission label.
* Patients with complete demographic, diagnosis, and medication details after preprocessing.

**Exclusions:**

* Weight column (~97% missing).
* Constant-value drug columns (examide, acetohexamide, glimepiride-pioglitazone).
* Duplicate encounter records.

**Data Sources:**

* **Primary:** UCI Machine Learning Repository – “Diabetes 130-US hospitals for years 1999–2008 Data Set.”
* **Supporting:** ICD mapping reference file for grouping diagnosis codes.

**Audience Level:**

* Analysis conducted at the **encounter level** (each hospital visit treated as a separate case).

**Variable Selection:**

* **Demographic:** Age (numeric midpoint), gender, race.
* **Utilization:** Number of inpatient, outpatient, and emergency visits in the past year.
* **Clinical:** Primary and secondary diagnoses (Diag\_1, Diag\_2, Diag\_3 grouped into clinical categories).
* **Lab Results:** A1C result, glucose serum levels.
* **Medications:** 23 anti-diabetic drug indicators, including indicators for change in medication.
* **Administrative:** Admission type, admission source, discharge disposition, payer code.
* **Derived Variables:** Missing indicators for payer, medical specialty, diagnosis, and weight; clinically relevant interaction features.

**Assumptions & Data Limitations:**

* Patients may have been readmitted to other hospitals not captured in this dataset.
* Missing values are either unrecorded or not applicable, and missingness itself may carry predictive information.
* The dataset reflects care patterns from 1999–2008 and may differ from current clinical practices.

## **4.0 Dependencies and Risks**

|  |  |
| --- | --- |
|  | Identification of key factors that may influence the outcome of the project and likelihood of it happening: |

|  |  |  |  |
| --- | --- | --- | --- |
| Risk | Likelihood | Impact | Mitigation |
| **High Class Imbalance** – Fewer <30-day readmission cases compared to non-readmissions | Medium | High | Applied multiple balancing methods (SMOTE, ROS, RUS, Tomek Links) on the **training set only** to improve recall and avoid leakage. |
| **Outlier Influence on Models** – Extreme values in numeric features may distort model training | Medium | Medium | Skewness assessment, Log2 transformation where effective, and IQR-based capping of outliers. |
| **Missing Values in Key Variables** – e.g., *Medical\_specialty*, *Payer\_code*, diagnosis fields | Medium | Medium | Imputed with “Missing”/“Unknown” categories; created missing indicators to retain predictive signal. |
| **Class Prior Shift** – Historical data (1999–2008) may not reflect current patient mix and practices | Medium | Medium | Recommend recalibration and monitoring if deployed in a modern setting. |
| **Data Quality Issues** – Potential coding errors or unrecorded follow-ups | Medium | Low–Medium | Careful variable grouping and imputation; exclude unreliable variables where necessary. |

## **5.0 Deliverable Timelines**

|  |  |
| --- | --- |
|  | List key dates and timelines as a work-back schedule. Activate line items based on complexity and line-of-sight required. Will set the stakeholder expectations for the process. |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Item | Major Events / Milestones | Description | Scope | Days | Date |
| 1 | Kick-off / Formal Request | Confirm dataset, define objective, and submit analytics plan after initial exploration | Project Kick-off | 3 | 14 Jul 2025 |
| 2 | Assessment / Triage | Explore structure and completeness of data, assess missing values, and map codes | Data Understanding | 3 | 17 Jul 2025 |
| 3 | Prioritization | Finalize relevant columns, drop/group irrelevant ones, handle duplicates, and set modeling direction | Variable Selection | 2 | 19 Jul 2025 |
| 4 | Data Exploration & Analysis | Conduct EDA to understand patterns in readmission, visualize demographic and treatment trends, group diagnosis codes, analyze repeat patients, and explore early readmission behavior | EDA & Insights | 4 | 23 Jul 2025 |
| 5 | Story Board 1 | Draft preliminary insights and visualizations, outline presentation story flow, and prepare early slides for review | Insight Framing | 2 | 25 Jul 2025 |
| 6 | QA Output | Build and evaluate models (Decision Tree, Random Forest, Logistic Regression, XGBoost, Neural Network), tune features, check data integrity, and prepare output for presentation | QA | 5 | 30 Jul 2025 |
| 7 | Internal Team Presentation | Share draft outputs and slides with advisor or peers, gather feedback on messaging, clarity, and insight delivery | Review Round 1 | 2 | 1 Aug 2025 |
| 8 | Go/No-Go | Decide if current model and insights are strong enough to move forward to final prep or require changes | Modeling Strategy | 1 | 2 Aug 2025 |
| 9 | Story Board 2 | Finalize charts, simplify visuals, update slide deck with improved storyline, ensure alignment with objectives and feedback | Final Prep | 4 | 6 Aug 2025 |
| 10 | Pilot | Conduct full rehearsals, test timing and flow, resolve any unclear messaging, and polish delivery | Rehearsal | 2 | 8 Aug 2025 |
| 11 | Delivery & Sign-off | Deliver final presentation to stakeholders, address any last-minute questions, and secure approval | Final Output | 3 | **12 Aug 2025** |