End to End Healthcare Analytics using Snowflake and AWS - 1

Business Overview

Healthcare analytics is the process of using data and analytical methods to improve the delivery of healthcare services and patient outcomes. One important area of healthcare analytics is the analysis of patient length of stay (LOS), which refers to the amount of time a patient spends in a healthcare facility.

Length of stay is a critical metric in healthcare, as it can impact patient outcomes, healthcare costs, and hospital capacity. By analyzing patient LOS data, healthcare providers can identify opportunities to improve the delivery of care and reduce costs.

In addition to improving patient outcomes, the analysis of patient LOS can also help healthcare providers to reduce costs. For example, by identifying patients who are at risk of extended LOS, providers can take proactive steps to ensure they receive the care they need in a timely manner.

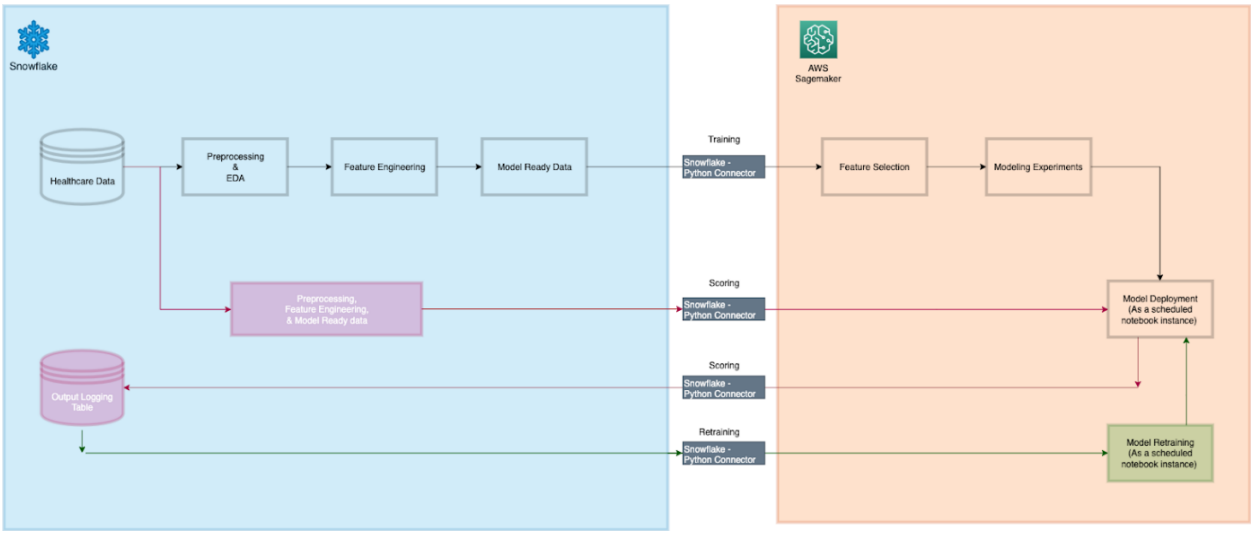
Overall, the analysis of patient length of stay is a valuable area of healthcare analytics that can help providers to improve patient outcomes and reduce costs. By leveraging data and advanced analytical techniques, healthcare providers can gain a deeper understanding of patient needs and identify opportunities to improve the delivery of care.

Approach

1. Introduction to Snowflake and Snowflake Worksheet
2. EDA in Snowflake
3. Feature Engineering in Snowflake
4. AWS Sagemaker Setup
5. Fetching the data from Snowflake using snowflake-connector-python, and snowflake-sqlalchemy
6. Data Preprocessing
7. Feature Selection
8. Model Building
    a. Linear Regression
    b. Random Forest Regression

    c. XGBoost Regression
9. Model Predictions
10. Inserting model predictions in Snowflake
11. Scoring function deployment and scheduling
12. Sending Status mail

Architecture



Aim

1. Data Analysis in Snowflake
2. Build a Machine Learning model to predict the length of stay for patients
3. To schedule the AWS Sagemaker Notebook
4. To perform live data scoring and inserting predictions to Snowflake
5. send status mail

Data

The training data is present in snowflake for about 230k patients across various regions and hospitals. There are total 19 features available in the data.
The simulation data is available for 71K patients for prediction purpose.

Tech Stack

- ➢ Tools: AWS Sagemaker, Snowflake
- ➢ Language: Python

➢ Libraries: snowflake-connector-python, snowflake-sqlalchemy, xgboost, pandas, numpy, scikit-learn

Prerequisites

- Snowflake Account
  - Snowflake Real Time Data Warehouse Project for Beginners-1
- AWS Account
- Understanding of basic SQL queries
  - SQL Project for Data Analysis using Oracle Database-Part 3
  - SQL Project for Data Analysis using Oracle Database-Part 4

Code Overview

The code files are available in the code.zip file. Once you unzip the code.zip file, you can find the following folders within it.

1. Data
2. Python files
3. SQL Queries

1. The Data folder contains health_data.csv and simulation_data.csv files.
   Note: It is assumed that the data is present in the snowflake table.
   You can refer this project to understand how to insert the data in snowflake.

2. The python files folder contains all the files from the jupyter environment created in AWS Sagemaker.

3. The SQL files folder contains all the SQL queries used in the snowflake worksheet.
   Note: If you want to learn more about SQL please refer to the following project
   - SQL Project for Data Analysis using Oracle Database-Part 3
   - SQL Project for Data Analysis using Oracle Database-Part 4

Project Takeaways

1. Understand Snowflake UI
2. Data Analysis using Snowflake Worksheet
3. What is CTE(common table expression) in SQL?
4. Feature Engineering in Snowflake

5. Fetching data from snowflake to python using the SnowSQL connector
6. How to set up Jupyter on AWS Sagemaker?
7. How to select the appropriate machine in AWS Sagemaker?
8. Data Preprocessing
9. How to select important features for model building?
10. Building and Predicting XGBoost model for LOS
11. Scheduling AWS Sagemaker Notebook
12. Sending email notifications for scheduling and tasks
13. Inserting the Predictions in Snowflake