Project Documentation

Project Title: Medical Appointment No-Show Prediction

Business Problem

A significant issue in medical setting is patients failing to attend scheduled doctor appointments despite receiving instructions (no-shows). Our client, a medical ERP solutions provider, seeks to tackle this by introducing a machine learning model into their software. This model aims to predict patient attendance, enabling medical providers to optimize appointment management.

Dataset Description

The dataset utilized in this project comprises appointment records from medical institutions, capturing various attributes related to patients and their appointments. Key features include:

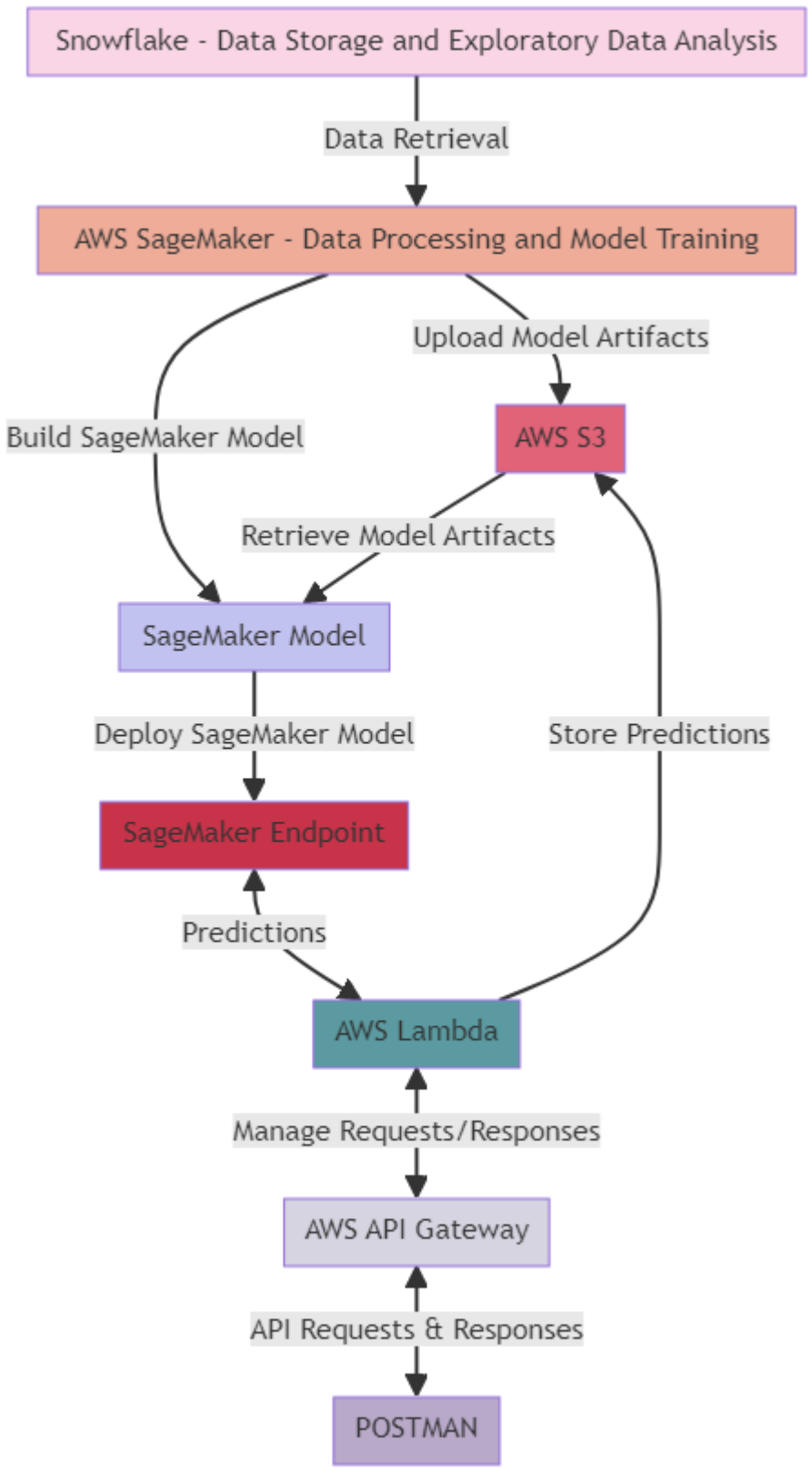
- **Patient demographics**: age and gender
- **Health characteristics**: the presence of conditions such as diabetes or hypertension
- **Appointment-specific details**: scheduled and appointment dates, and whether the patient received a reminder SMS
- **Target**: binary indicator representing whether a patient was a no-show or attended their appointment.

No	Column Name	Description
01	PatientId	Identification of a patient
02	AppointmentID	Identification of each appointment
03	Gender	Male or Female. Female is the greater proportion, women take way more care of their health in comparison to men.
04	ScheduledDay	The day someone called or registered the appointment, this is before the appointment of course.
05	AppointmentDay	The day of the actual appointment, when they have to visit the doctor.
06	Age	How old is the patient.
07	Neighbourhood	Where the appointment takes place.
08	Scholarship	True or False. Indicates whether the patient is enrolled in Brasilian welfare program Bolsa Família.
09	Hipertension	True or False. Indicates if the patient has hypertension.
10	Diabetes	True or False. Indicates if the patient has diabetes.
11	Alcoholism	True or False. Indicates if the patient is an alcoholic.
12	Handcap	True or False. Indicates if the patient is handicapped.
13	SMS_received	True or False. Indicates if 1 or more messages sent to the patient.
14	No-show	True or False (**Target variable**). Indicates if the patient missed their appointment.

This rich dataset provides a comprehensive view of factors potentially influencing patient attendance, enabling the development of a nuanced predictive model.

Solution approach

- **Model Development:** Created a machine learning model to assess the likelihood of patient no-shows, enhancing appointment scheduling efficiency.
- **System Integration:** Deployed the model with an API for integration into the client's ERP system, this allows real-time predictions, streamlining the ERP's existing workflow.



Data Loading

The initial phase of the project involved establishing a robust data infrastructure for effective management and access to the dataset. This process encompassed two key phases:

- **Snowflake Database Creation**: A database was created in Snowflake, a cloud-based data warehousing platform, specifically designed to store and manage the appointment records. This setup ensured secure and scalable handling of the data.
- **Data Retrieval in AWS SageMaker**: With the Snowflake database in place, AWS SageMaker, a cloud machine learning platform, was utilized to connect to the database. This connection facilitated seamless retrieval of the dataset for further processing. SageMaker provided the computational power and tools necessary for handling data loading, preprocessing, and model development.

This approach to data loading ensured a streamlined workflow, laying a solid foundation for the subsequent stages of data preprocessing and feature engineering.

Data Preprocessing and Feature Engineering

In preparing the dataset for modeling, a series of data preprocessing and feature engineering steps were undertaken:

- **Data Cleaning**: Involved handling missing values, removing records with unrealistic values (e.g., negative ages), and dropping records where the 'AppointmentDay' was earlier than the 'ScheduledDay'.
- **Feature Transformation**: Included transforming 'AppointmentDay' and 'ScheduledDay' to calculate the time interval between scheduling and the actual appointment, and categorizing age into meaningful groups.
- **Feature Encoding**: The 'Gender' attribute was encoded, and the 'Neighbourhood' feature underwent target encoding.

Additionally, to address the challenge of class imbalance in the dataset, several specialized datasets were created:

- **Upsampled Dataset**: Increased the representation of the minority class by duplicating observations.
- **Downsampled Dataset**: Reduced the size of the majority class by randomly removing observations.
- **SMOTE Dataset**: Utilized Synthetic Minority Over-sampling Technique to generate synthetic samples for the minority class.

These steps ensured the dataset was not only clean and consistent but also balanced and enriched with derived features, providing a robust foundation for the predictive model.

Modeling Approach

The modeling process was methodically structured to ensure the selection of the most predictive features and the most effective model:

1. **Feature Correlation Analysis**: Prior to feature selection, a correlation analysis was conducted to identify and understand the relationships between different features. This helped in recognizing any multicollinearity issues and informed the decision on which features to retain or modify.
2. **Feature Selection**: Feature importance was assessed using Logistic Regression and Decision Tree models. Features with at least 1% importance in either model were retained, leveraging the union of important features from both models for a balanced approach.
3. **Model Selection and Dataset Evaluation**: Four algorithms – Logistic Regression, Decision Tree, Random Forest, and XGBoost – were compared using ROC AUC and F1 score metrics. XGBoost, trained on the original dataset, emerged as the best performer.
4. **Feature Importance Reassessment**: Post model selection, feature importance was reassessed with XGBoost, further refining the feature set for optimal impact.
5. **Hyperparameter Tuning**: Hyperparameters were fine-tuned using Hyperopt, exploring a wide range of values to enhance the model's accuracy.
6. **Final Model Training**: The final model, trained with the optimized hyperparameters, showed improved ROC AUC and F1 scores, indicating enhanced predictive accuracy.

Model Deployment

This section outlines the deployment process of the XGBoost model, focusing on its integration into the AWS cloud infrastructure. Key steps in the deployment process include:

1. **Model Artifacts Preparation**: The trained XGBoost model and necessary scripts were packaged into model artifacts, ready for deployment.
2. **Uploading to AWS S3**: These artifacts were then uploaded to an Amazon S3 bucket, ensuring their availability for AWS services.
3. **SageMaker Model Creation**: A SageMaker model was created, specifying the S3 location of the model artifacts and configuring the appropriate runtime environment.
4. **Endpoint Deployment**: The model was deployed to a SageMaker endpoint, enabling it to provide real-time predictions.
5. **Lambda Function for Model Invocation**: An AWS Lambda function was developed to act as an intermediary between the SageMaker endpoint and the API Gateway, managing HTTP requests and model responses, sending the predictions back to AWS S3 for storage.
6. **API Gateway for REST API**: An Amazon API Gateway was set up to expose a RESTful API, making the model's predictions accessible for external applications.
7. **Testing with Postman**: The final step involved rigorously testing the deployed model using Postman to ensure its functionality and reliability.

ERP Integration

In theory, integrating this model into an ERP system would involve establishing a connection between the ERP software and the AWS-hosted model endpoint. This could be achieved through API calls from the ERP system to the AWS API Gateway, which would then relay predictions from the model back to the ERP system. Such integration would enable the ERP software to utilize the

predictive model for enhancing appointment management, providing real-time insights and recommendations based on the model's predictions.

Business Impact and Future Prospects

1. **Utility for Stakeholders**: This solution significantly benefits healthcare providers, administrative staff, and patients by predicting appointment no-shows. It enables more efficient scheduling, reduces waiting times, and optimizes resource utilization. The model's insights can lead to cost savings and improvements in patient care and satisfaction, enhancing the overall healthcare delivery experience.
2. **Business Impacts**: The deployment of this predictive model can lead to increased operational efficiency and financial savings for healthcare facilities. By reducing no-show rates, healthcare providers can better manage their schedules and resources, potentially increasing revenue and improving service quality. The model's data-driven approach also offers valuable insights for strategic decision-making and policy development within healthcare organizations.
3. **Future Prospects**: The model presents opportunities for further development and scalability. Future enhancements could include integrating with broader healthcare IT systems, refining predictions with more extensive datasets, and adapting the model to other critical healthcare outcomes. The evolving landscape of machine learning and healthcare technology promises to expand the model's capabilities and applications, ensuring its long-term relevance and effectiveness in the healthcare sector.