Explanatory document of the prototype

📊 Hotel Booking Cancellation Predictor

This is a **prototype dashboard** designed for hotels to **track new bookings** in real time and visualize their details. The system **predicts the likelihood of each booking being canceled** based on historical data. Additionally, users can view **recent bookings and their associated information**.

Use:

```
streamlit run appv3.py
```

Which Model Did You Choose, From Which Course, and Why?

The dataset used in this project is a **large hotel booking dataset**, containing extensive historical data on hotel reservations, including whether each reservation was ultimately canceled or not.

The approach taken was to develop a **simple logistic regression model** to predict whether new bookings **are likely to be canceled** based on past data.

This dataset was previously used in the **third assignment of the "Cloud Computing" course** during the **1st Term**, where we deployed an AWS-based booking cancellation prediction API using **Lambda, API Gateway, and S3**. At that time, we used a very basic model with only two selected features from the dataset. The dataset caught my attention due to its **large size and rich set of features**, so I wanted to explore it in more depth.

What Is the Purpose of This Prototype?

In a **real-world scenario**, this prototype could be a valuable tool for **hotel sales and marketing teams**. It would allow them to **monitor new bookings in real time** and get **immediate predictions** on whether those bookings might be canceled, enabling them to take proactive measures.

What Were the Main Challenges?

The biggest challenge was **preprocessing such a large dataset**, both in terms of **rows (volume)** and **columns (features)**. It required a thorough **feature selection process** and **careful data treatment** to optimize model efficiency.

Additionally, the **large number of observations** made **model selection complex**. My initial goal was to use **GridSearchCV** to find the best model and hyperparameters, but due to the computational cost, this was **not feasible within a reasonable timeframe**. Addressing this issue is a **top priority for future developments**.