Big Data Analytics Project

May 2025

San Francisco Crime Route Analyser

Amber Gonzalez-Pacheco
Karishma Mehta
Shivani Tayade
Sakshi Kumar

Business Objective

San Francisco’s overall reported crime rate recently made headlines as it fell to its lowest level since 2001, according to SFPD data released by former Mayor London Breed. Property crime dropped 31%, and violent crime by 14%, between 2023 and 2024. Human trafficking declined 45%, while robberies fell 22%, burglaries were down 11%, and assaults dropped by 6%. However, despite the improvements, public perception of “out-of-control” crime remained a decisive issue in the 2024 mayoral race, signalling a gap between data and residents’ sense of safety. This disconnect means residents remain wary when walking in certain neighbourhoods or late-night hours, and through first-hand experiences living in the city, the team felt the need to bridge this gap by predicting block-level crime risk, and guiding San Francisco pedestrians towards safer routing decisions along walking paths in real time.

Key Actionable Business Initiative

To address the gap between declining crime statistics and persistent safety concerns among San Francisco pedestrians, our team implemented a Streamlit-based chatbot that delivers real-time, low-risk walking route recommendations. The primary aim of this initiative was to operationalize predictive analytics, enabling end-users to make safer travel decisions with minimal effort.

The chatbot operates by accepting user-provided origins, destinations, and intended travel days of the week and times. It then systematically evaluates multiple possible walking routes, leveraging a custom risk scoring model built on recent and historical crime data. This model incorporates variables such as time of day, crime type, and geographic location at the block level, calculating and assigning a relative risk score to each potential path. Where elevated risks are detected, the backend algorithm recomputes the next safest route with an average risk score below the threshold, and in turn the chatbot recommends alternative routes that minimize exposure to those risk factors.

To maximize accessibility and ease of use, the application features a streamlined interface compatible with both desktop and mobile browsers. The system architecture integrates real-time Kafka streaming to ensure rapid processing of incoming data and prompt response times, enabling a live feed of crime incidents to enhance users’ awareness before making their journey. During the pilot phase, the chatbot was embedded on a local host website powered by Streamlit, and randomized dummy crime incidents were generated to provide live updates in the feed, thus facilitating iterative feedback and simulating a realistic usage scenario.

This initiative translates advanced predictive analytics into actionable, context-sensitive guidance for everyday users. By integrating a real-time risk scoring model with a responsive chatbot interface, the solution makes data-driven safety insights directly accessible, empowering San Francisco pedestrians to make more informed and safer routing decisions.

Metrics of Success

To evaluate the success of our crime-aware routing initiative, we focus on three key metrics that aligned with our goals of maximizing user safety, system responsiveness, and adoption:

First, the most important metric is the Route Risk Reduction Rate, which measures the average decrease in predicted crime risk when users are rerouted from their original path. Our goal is to achieve at least reduction in route risk to below 0.5 by intelligently avoiding the riskiest areas, identified through our model's spatial risk predictions.

The second metric is High-Risk Route Flagging Recall, which evaluates how effectively the system identifies all truly unsafe routes. In a public safety context, we deliberately prioritize recall over precision to ensure that we capture as many dangerous segments as possible, even if it means occasionally flagging some safe ones. This trade-off reflects a critical design decision: it's better to over-warn than to risk underestimating potential threats.

Lastly, System Responsiveness remains important for user adoption. We target a median response time of under 3 seconds, enabling real-time interaction via the Kafka-streamed backend and Streamlit interface. Together, these metrics define the impact, reliability, and user trust in our platform.

Role of Analytics

Analytics adds value to this initiative in three specific and interdependent ways: enablement, refinement, and evaluation.

First, analytics enables the entire functionality of the application, without the predictive model trained on historical San Francisco crime data, the chatbot would not be able to assign risk scores to each walking route. The Random Forest classifier transforms location, time of day, and day of week into a probability that a pedestrian might encounter a high-risk incident, allowing the system to operationalize crime data into real-time safety decisions.

Second, analytics refined the business logic and design through iterative insight. For example, exploratory analysis showed that crimes were overreported at common timestamps like 00 and 30 minutes past the hour, prompting us to bucket time features into 15-minute intervals to improve model accuracy. Predictive model performance also guided decisions such as excluding redundant neighborhood data once latitude and longitude proved more predictive. The rerouting logic itself evolved based on analytics feedback, moving from static avoidance zones to multi-pass rerouting with dynamic risk minimization.

Finally, analytics serves as the basis for evaluating success. We prioritized model recall to reduce the chance of underestimating dangerous routes, and we monitored the average risk score of routes before and after rerouting to validate effectiveness. Metrics such as reduced high-risk route frequency, improved user path safety, and responsiveness to recent crimes all stem from the analytical design of the system. Analytics not only powers the backend, it informs the project’s strategy, architecture, and real-world impact.

Thinking Through the Analytics

For this project, we relied on publicly available data from the DataSF Open Data Portal, specifically the San Francisco Police Department’s crime incident reports from 2018 to the present. While the original dataset spanned over seven years, we found that including all years introduced noise and diluted the predictive power of our model. After exploratory analysis, we truncated the dataset to only the most recent three years, which provided more meaningful patterns and better aligned with our goal of near-real-time risk forecasting.

Our outcome variable is a probabilistic crime risk to pedestrian score at each GPS coordinate in San Francisco, aggregated at the 15-minute interval level. The explanatory features include latitude, longitude, hour and minute of day, day of week (one-hot encoded), and the severity of the crime. Initially, we considered using neighborhood labels as a feature but found them redundant due to the granularity of latitude and longitude. To model this, we used a class-weighted Random Forest Classifier trained to predict the probability of a crime occurring at a specific coordinate and time. This predictive setup allowed us to not only score individual routes but to intelligently suggest reroutes that avoid high-risk zones.

The variation in our dataset comes from both spatial and temporal dimensions, as well as the severity and frequency of reported crimes. Crime types ranged from low-severity incidents like traffic violations and gambling to high-severity events like homicide and rape. Since these categories have vastly different implications for pedestrian safety, we introduced a weighting scheme that prioritized recency and severity, ensuring that rare but dangerous events are appropriately represented in the model’s risk assessment.

This project is primarily predictive analytics, though exploratory analysis was critical in guiding feature selection and understanding crime patterns. Our aim was not just to predict with high accuracy, but to provide actionable insights for users seeking safer walking routes in real time. As such, model interpretability and usability were just as important as raw predictive performance. Streamlit interfaces and Kafka-powered pipelines helped convert predictions into practical safety decisions.

Impediments we faced included inflated risk scores in historically dangerous locations, where outdated but frequent crime reports created a false sense of persistent danger. To resolve this, we adjusted the weighting of crime incidents to prioritize recency and assign higher weights to crimes that pose more immediate threats to pedestrians, such as robbery or assault. This allowed our model to more accurately reflect the present-day risk environment and better serve the needs of users navigating the city on foot.

One of the biggest challenges we faced during this project was the artificial clustering of crime reports around the top and bottom of the hour. Many incidents were logged at rounded times like 12:00 or 12:30, not because that's when they occurred, but likely due to estimation or reporting defaults. This led to misleading spikes in risk at those intervals, which distorted the model's understanding of actual crime patterns. To address this, we implemented time bucketing and smoothing strategies to distribute risk more evenly and avoid overemphasizing these artificially inflated time slots, resulting in more accurate and actionable route scoring.

Executing the Analytics

Our analytics pipeline was developed and executed by a small team of data science practitioners responsible for the entire end-to-end workflow. Data collection was handled through automated ingestion from the DataSF portal, followed by preprocessing, feature engineering, and model training. Model development, evaluation, and tuning were conducted iteratively with ongoing testing via Jupyter Notebooks and validation pipelines. Once the model was finalized, we implemented it in a real-time Kafka streaming architecture, enabling live scoring and rerouting functionality within a Streamlit user interface.

Although this project was largely self-contained, all team members contributed to defining success metrics, assessing trade-offs between precision and recall, and evaluating rerouting strategies based on predictive risk. The integration of modeling and deployment responsibilities helped ensure alignment between technical goals and user experience. In a future production setting, responsibilities could be distributed across data engineers (for ingestion), machine learning engineers (for modeling and monitoring), and product teams (for UI integration and metrics evaluation). For now, a centralized approach helped ensure a streamlined and cohesive build-out of the analytics-driven safety application.

Implementation and Scale

Once the route risk scoring results are available, they directly influence the routes recommended to users. Rather than offering static directions, the system dynamically adjusts suggested paths based on real-time past 24 hour crime predictions. This allows pedestrians to avoid high-risk areas, particularly at vulnerable hours or in volatile neighborhoods. The project has already informed critical decisions, like rerouting logic, severity-based crime weighting, and real-time Kafka streaming, all of which would not exist without the analytics layer. We plan to embed the model's predictions directly into a user-friendly chatbot interface (Streamlit) that mimics existing mapping apps, lowering the barrier for adoption.

In terms of scaling, organizational challenges could include data freshness, user trust, and technical limitations. Relying on up-to-date crime feeds (currently simulated) will require official API access or partnerships with city data providers. Technically, expanding from a local Streamlit app to a fully deployed mobile platform will require reengineering for scalability and responsiveness. There are also cultural hurdles, users must trust that the model makes valid, unbiased recommendations, which means prioritizing transparency and fairness in model design. To address these, we plan to transition from a prototype to a production-ready system by incorporating continuous feedback loops, retraining the model monthly, and updating the crime feed to reflect more recent or real-time incidents. Our next step is meeting with the team at Google Maps to embed our idea into the walking option as a beta feature for San Francisco users to choose, eventually broadening the crime data to all other major cities where it is accessible. This ensures that our analytics are not a one-time solution but an evolving tool embedded in a broader safety ecosystem for urban mobility.