Final Assignment: Applying Data Science Methodology to Hospitals

CRISP-DM Data Science Methodology Project

Topic Chosen

Hospitals

1. Business Understanding

Client Role - Business Problem:

A large hospital network is struggling with emergency room (ER) overcrowding and long patient wait times. These issues are leading to decreased patient satisfaction and increased operational stress. Hospital administrators want to use data science to predict ER patient volume ahead of time so that staff can be scheduled appropriately and resources can be managed efficiently.

Business Objective:

Develop a predictive model that can forecast ER visits 1 day in advance using historical patient and hospital data.

Success Criteria:

- Predict patient volume with at least 85% accuracy
- Enable hospital staff to plan shifts and allocate beds more efficiently
- Reduce average ER wait time by 20%

2. Data Understanding

Data to Be Collected:

- Historical ER visit records (date, time, reason, triage level)
- Weather data (temperature, rain, holidays)
- Patient demographics (age, gender)
- Seasonal and event data (e.g., flu season, festivals)

- Hospital resource data (staff per shift, bed occupancy)

Exploratory Data Analysis:

- Determine patterns in visits by hour, day, and month
- Visualize visit trends (e.g., spikes during flu season or weekends)
- Identify anomalies or outliers in the data (e.g., sudden spikes)

3. Data Preparation

Steps Taken:

- Clean data: Handle missing values in temperature, triage categories, and patient records
- Normalize: Scale numerical features like temperature and staff count
- Categorize: Convert day of the week, holiday indicator, and weather conditions to categorical variables
- Feature Engineering:
  - is_weekend (Boolean)
  - previous_day_visits
  - flu_alert_level
  - staff_to_bed_ratio

4. Modeling

Goal: Predict the number of ER visits for the next day.

Models Considered:

- Linear Regression - as a baseline
- Random Forest Regressor - for more accuracy with non-linear relationships
- Gradient Boosted Trees - for performance optimization

Training Approach:

- Train on 80% of the data

- Test on the remaining 20%

- Use cross-validation to check model consistency

5. Evaluation

Evaluation Metrics:

- Root Mean Square Error (RMSE)

- Mean Absolute Error (MAE)

- R^2 Score

Results:

- Best model: Gradient Boosted Trees

- R^2 Score: 0.89

- RMSE: Within acceptable error margin (±12 patients)

- Model achieves business goal of >=85% accuracy

6. Deployment

Deployment Plan:

- Integrate model into the hospital's daily operations dashboard

- Automatically pull data every night to generate next-day forecasts

- Alert ER managers via email and internal system about expected volume

User Interface:

- Simple web dashboard showing:

  - Predicted patient volume

  - Suggested staffing level

  - Historical vs. predicted trend lines

7. Feedback

Continuous Improvement Strategy:

- Weekly feedback from ER supervisors on forecast accuracy
- Track actual vs. predicted values to detect model drift
- Monthly retraining using most recent data
- Adjust features based on feedback (e.g., include festival data)

Summary

CRISP-DM Stage	Task Completed
Business Understanding	Defined hospital's ER overcrowding problem
Data Understanding	Identified and explored relevant datasets
Data Preparation	Cleaned and engineered features for modeling
Modeling	Trained multiple models and selected the best one
Evaluation	Evaluated using R^2 and RMSE; model met business goals
Deployment	Planned integration with hospital workflow
Feedback	Designed continuous feedback loop with review plan