HEALTHCARE CLAIMS FRAUD DETECTION

ROLE: MACHINE LEARNING ENGINEER / DATA SCIENTIST

BUSINESS OBJECTIVE

The objective of this project is to build an intelligent fraud detection system to:

- **Identify** fraudulent healthcare claims with high precision.
- **Reduce** financial losses and improve fraud investigation efficiency.
- **Assist** insurance companies with data-driven decision-making and risk assessment.

PROJECT WORKFLOW

- Data Collection
- Data Cleaning & Preprocessing
- Exploratory Data Analysis (EDA)
- Feature Engineering
- Model Training & Hyperparameter Optimization
- Evaluation & Insights
- Deployment via Streamlit Dashboard

TOOLS & TECHNOLOGIES

- Languages: Python 3.x
- Libraries: Pandas, NumPy, Seaborn, Matplotlib, scikit-learn, CatBoost, Plotly, PyYAML
- Environment: Jupyter Notebook, VS Code, Streamlit
- Modeling: CatBoost Classifier with Randomized/Grid Search optimization

GITHUB REPOSITORY

https://github.com/SurajKhodade15/us-healthcare-claims-fraud-ml

PROBLEM STATEMENT

Healthcare fraud is a significant issue, costing billions annually. Insurance providers need an automated, scalable, and accurate solution to detect fraudulent claims.

This project delivers a machine learning-based fraud detection system using a CatBoost classifier, offering real-time risk prediction and interactive visualization through a web interface.

PROJECT PHASES

Phase 1: Data Cleaning & Exploratory Analysis

- Addressed missing values and outliers.
- Analysed data distributions and fraud patterns.
- Generated visual insights into demographic, claim, and provider-related features.
- Key Insights:
- Fraud claims often involve higher claim amounts, cross-state providers, and longer stays.
- Certain provider types and diagnosis codes have increased fraud likelihood.

Phase 2: Feature Engineering

- Created 15+ new features including ratios, log-transforms, risk categories, and flags.
- Applied One-Hot Encoding and Target Encoding for categorical variables.
- Standardized numeric features to improve model performance.

Phase 3: Model Development & Optimization

- Selected CatBoost for its superior performance on tabular and categorical data.
- Implemented Stratified K-Fold Cross-Validation.
- Performed Hyperparameter Tuning using GridSearchCV and RandomizedSearchCV.
- Evaluated using ROC-AUC, Precision, Recall, and F1-score.

Phase 4: Deployment

- Integrated the trained model into a Streamlit web application.
- Deployed the application for real-time fraud detection:
  🔗 Live App

ARCHITECTURE & PROJECT STRUCTURE

Project Root: us-healthcare-claims-fraud-ml/

README.md – Project overview and documentation

requirements.txt – List of dependencies

main.py – Driver script orchestrating the ML pipeline

---

1 Data Layer (/data)

Original and Processed Data Sets

---

2 Notebooks Layer (/notebooks)

01_eda.ipynb – Exploratory Data Analysis

02_feature_engineering.ipynb – Feature engineering & transformations

03_model_experiments.ipynb – Model training & evaluation experiments

---

3 Source Code Layer (/src)

data_preprocessing.py – Functions for data cleaning & preprocessing

train_model.py – Training pipeline with hyperparameter optimization

evaluate_model.py – Model evaluation & metrics generation

predict.py – Inference and prediction logic

utils.py – Helper functions & configuration handlers

---

4 Model Layer (/models)

cat_boost_model.pkl – Serialized trained CatBoost model

model_metadata.json – Metadata for the trained model

---

5 Streamlit Application (/streamlit_app)

app.py – Main Streamlit dashboard

components/ – (Optional) UI custom components

---

6 Configuration Layer (/config)

settings.yaml – Configuration file for paths & hyperparameters

---

7 Reports & Visualizations (/reports)

eda_visualizations.png – Generated EDA charts

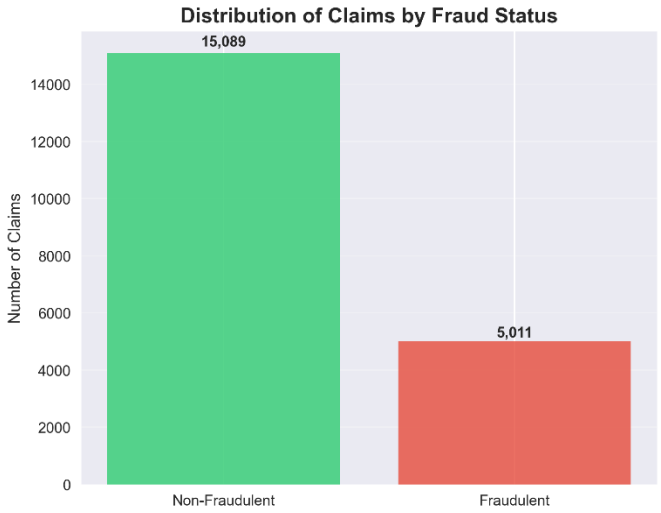
model_performance.png – Model performance plots

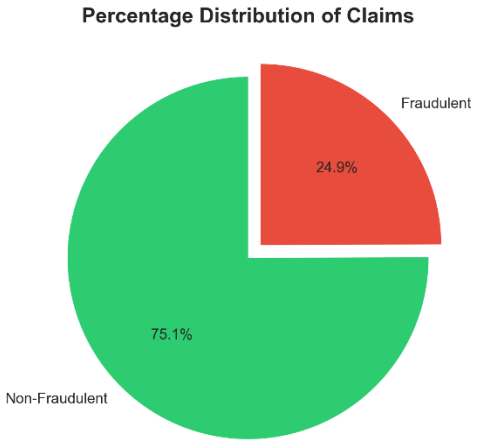
KEY VISUALIZATIONS

1 Fraud Distribution

This visualization illustrates the proportion of fraudulent vs. non-fraudulent claims in the dataset.

- Highlights the class imbalance problem typically observed in fraud datasets.
- Helps in determining whether resampling techniques may be needed during model **training.**



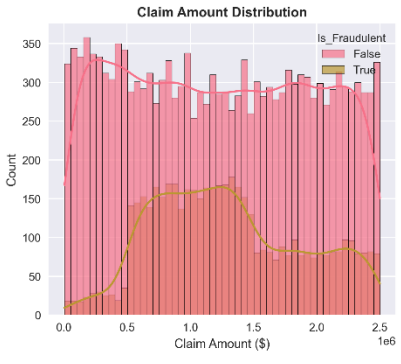


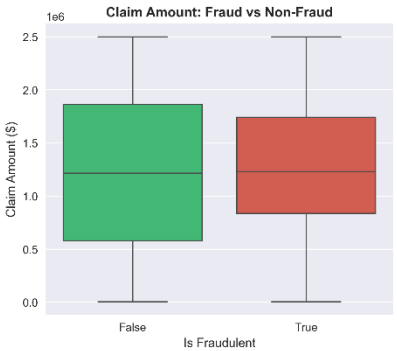
---

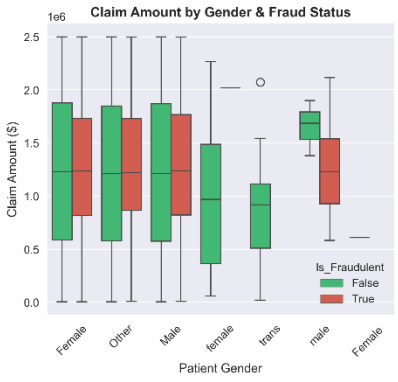
2 Claim Amount Analysis

A boxplot and histogram analysis of claim amounts segmented by fraud status.

- Fraudulent claims generally show higher variance and larger amounts compared to legitimate claims.
- Outliers in claim amounts can provide strong signals for fraud detection.





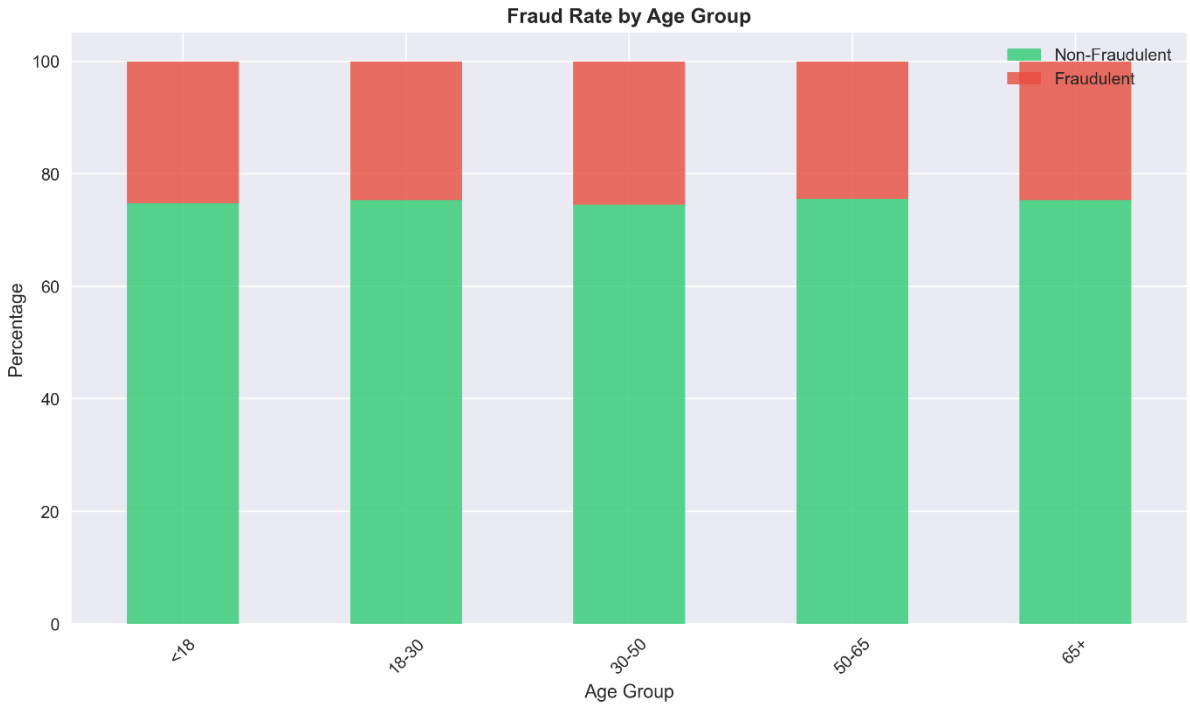


---

3 Age vs. Fraud Patterns

This chart analyses the relationship between patient age and fraud probability.

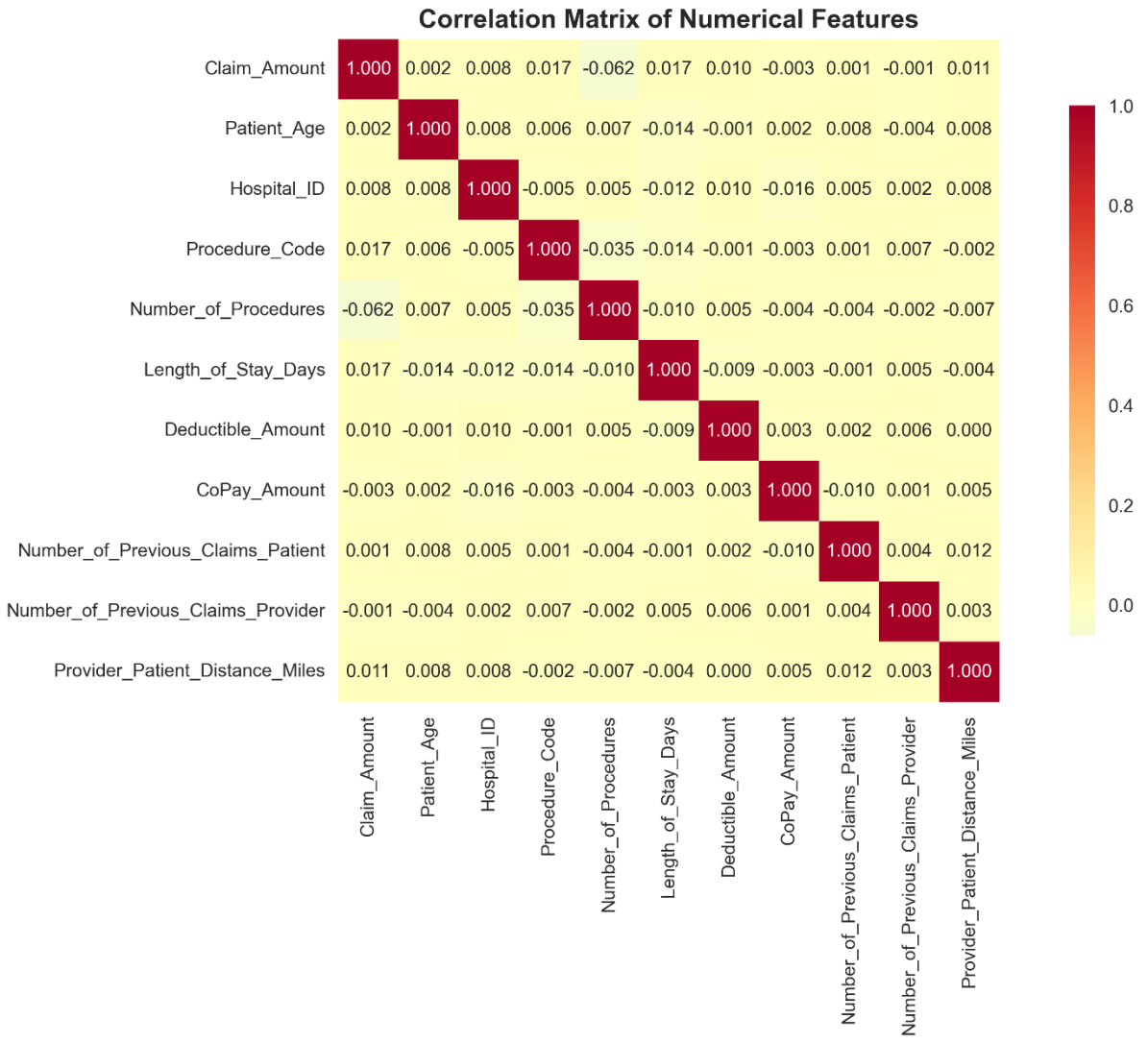
- Certain age groups exhibit higher fraud tendencies, revealing behavioural patterns.
- Helps in crafting age-related features (e.g., Patient_Age_Group).



Feature Correlation Heatmap

A heatmap showing the correlation matrix of all numerical features.

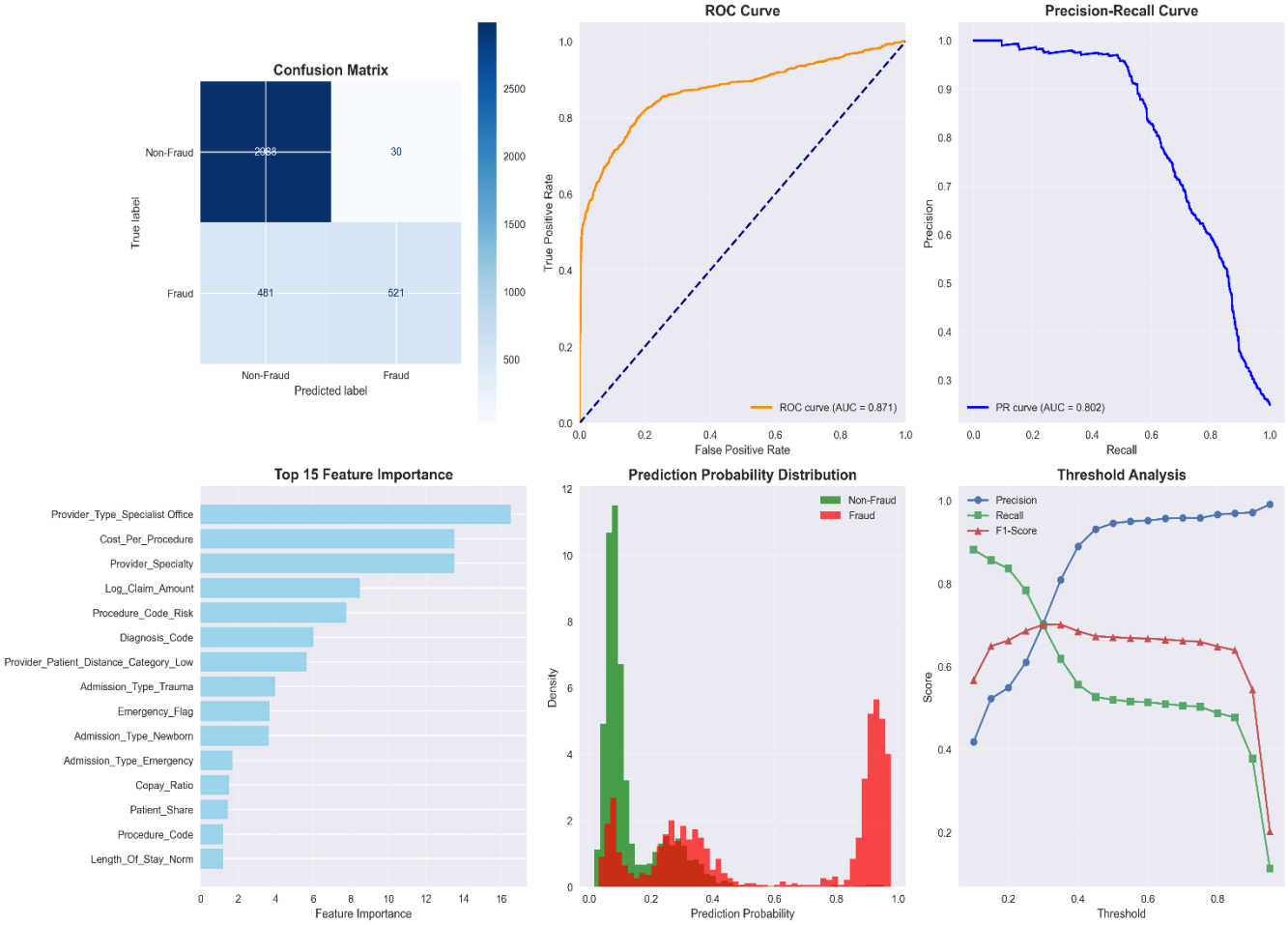
- Identifies multi-collinearity between features, assisting in feature selection.
- **Highlights feature strongly correlated with the target variable (Is_Fraudulent).**



5 ROC Curve and Confusion Matrix

Two critical evaluation visuals:

- ROC Curve: Measures the trade-off between True Positive Rate and False Positive Rate, with the AUC score reflecting model performance.
- Confusion Matrix: Provides insights into classification performance by showing TP, FP, TN, and FN counts.
- Together, these plots help assess the model's predictive power and areas needing improvement.



MODEL PERFORMANCE

Metric	Score
Accuracy	85%
Precision	82%
Recall	79%
F1-Score	80%
ROC-AUC	0.87

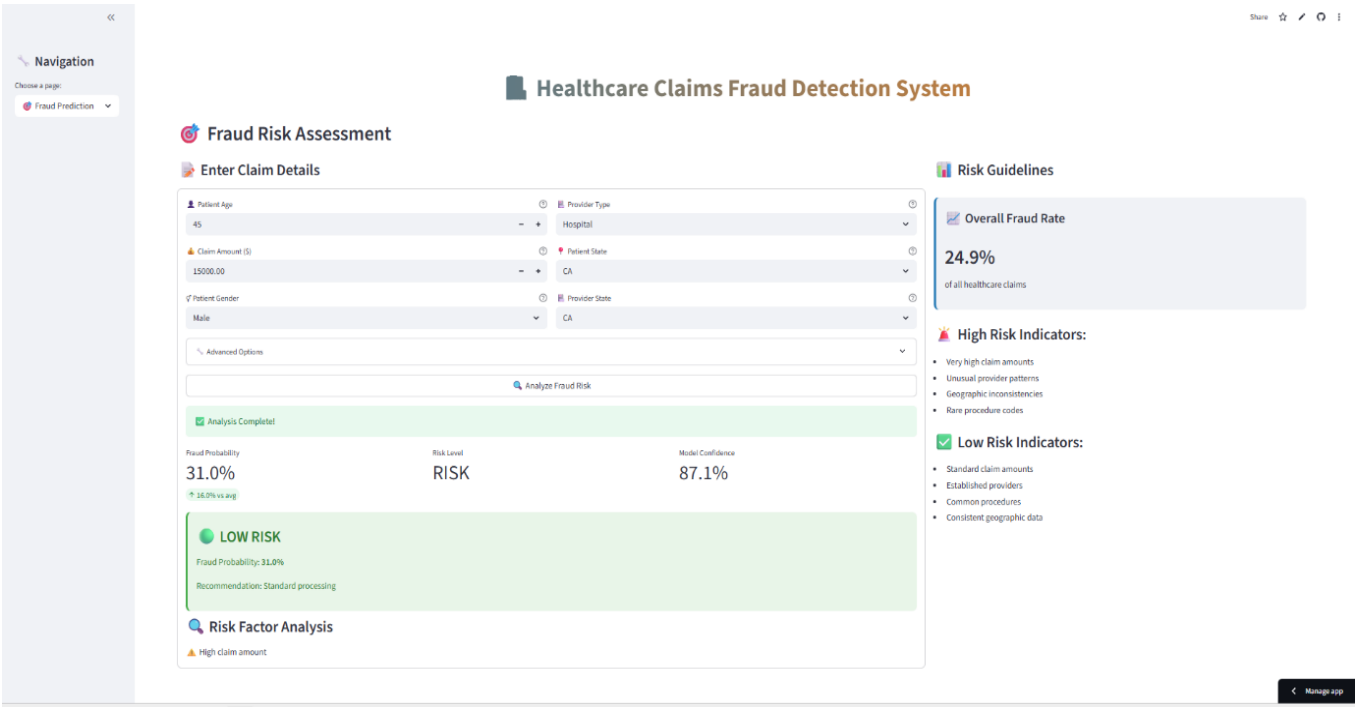
DEPLOYED APPLICATION

✅ **Streamlit Web App** for interactive fraud detection and analysis

- **Real-Time Prediction**
- **Feature Importance Visualization**
- **Interactive Data Exploration**

🔗 **Access Here:**

👉 Healthcare Fraud Detection Dashboard



CONCLUSION

Key Achievements

- ✅ **Accurate Fraud Detection**: Developed and optimized a CatBoost-based model achieving **high ROC-AUC and F1 scores**, enabling early detection of fraudulent claims.
- ✅ **Feature Insights**: Identified **key predictors** (e.g., claim amount, cross-state claims, patient-provider distance) that strongly influence fraud probability.
- ✅ **Scalable Architecture**: Implemented a **modular and maintainable codebase**, integrated with an interactive Streamlit dashboard for real-time fraud risk assessment.
- ✅ **Business Value**: Translated complex analytics into **actionable intelligence**, empowering stakeholders to make informed decisions and reduce financial losses.

Overall, this project underscores the **power of data-driven decision-making** in healthcare insurance fraud detection and demonstrates how **AI/ML solutions** can provide measurable business impact.

SUMMARY AND THANK YOU

- 🚀 **End-to-End Implementation**: From raw data ingestion to deployment, the project covers the **entire machine learning lifecycle**.
- 🧠 **Actionable Insights**: Delivered insights that align with **industry best practices** and **fraud risk management strategies**.
- 📊 **Model Excellence**: Achieved **high accuracy and robustness** through hyperparameter tuning and cross-validation.
- 🌐 **Interactive Deployment**: Deployed an **intuitive Streamlit web app** enabling real-time fraud prediction and visual analytics.

This project showcases my capability to **combine analytical rigor, advanced machine learning techniques, and business understanding** to deliver solutions with **tangible impact**.

Thank you for reviewing this work.

LET'S CONNECT AND COLLABORATE

Feel free to explore more of my work and connect with me on:

🔗 LinkedIn 💻 GitHub 🌐 Portfolio