Top 10 Data Science Projects

NumPy

PyTorch

scikit learn

pandas

TensorFlow

NLTK

Credit Card Fraud Detection

Goal: Identify fraudulent transactions from a dataset.

Why it matters: Imbalanced data, real-world stakes, classification under uncertainty.

Tech stack: Python, scikit-learn, XGBoost, SMOTE for oversampling.

Dataset: Kaggle – Credit Card Fraud Detection

Customer Churn Prediction

Goal: Predict whether a customer is likely to leave a service.

Why it matters: Common in telecom, SaaS, and banking. Involves feature engineering and classification.

Tech stack: Python, pandas, LightGBM, SHAP for interpretability.

Dataset: IBM Telco Dataset

Movie Recommendation System

Goal: Recommend movies based on user preferences or behavior.

Why it matters: Real-world applications in streaming and e-commerce. Teaches collaborative filtering, matrix factorization.

Tech stack: Python, Surprise, LightFM, or PySpark.

Dataset: MovieLens 100k/1M

Sentiment Analysis on Tweets

Goal: Classify tweet sentiment (positive, neutral, negative).

Why it matters: NLP 101 + Twitter data = highly relevant. Can expand to hate speech detection or political analysis.

Tech stack: Python, NLTK/spaCy, transformers, BERT.

Dataset: [Sentiment140 or Twitter API]

House Price Prediction

Goal: Predict house prices based on features like size, location, etc.

Why it matters: A regression classic. Great for feature engineering and model tuning.

Tech stack: Python, scikit-learn, XGBoost, EDA with Seaborn.

Dataset: Kaggle – Ames Housing

Time Series Forecasting

Goal: Forecast future values using past data.

Why it matters: Most real-world data is time-dependent understanding trends and seasonality is critical for smart forecasting.

Tech stack: Python, Prophet, ARIMA, LSTM.

Dataset: [Yahoo Finance via yfinance package]

Image Classification

Goal: Classify images into categories.

Why it matters: Deep learning basics. CNNs, transfer learning, overfitting it's all in here.

Tech stack: Python, TensorFlow/PyTorch, OpenCV.

Dataset: Kaggle – Dogs vs Cats

Resume Parser or Job Matcher

Goal: Extract structured info from resumes or match to job descriptions.

Why it matters: Applied NLP + text classification + regex + vector similarity.

Tech stack: spaCy, Python, scikit-learn, FAISS.

Dataset: Self-curated or scrape job boards and resumes.

Fake News Detection

Goal: Classify whether a piece of news is real or fake.

Why it matters: High relevance. Involves NLP, classification, and ethics.

Tech stack: Python, TF-IDF, LSTM, transformers.
Dataset: Kaggle – Fake News

EDA Dashboard

Goal: Build an interactive dashboard that visualizes key insights from a dataset.

Why it matters: Communication is half the job. EDA + storytelling = win.

Tech stack: Python, Plotly, Dash, Streamlit.
Dataset: Any interesting one COVID, Netflix, Spotify, etc.