BANK LOAN DEFAULT PREDICTION

USING MACHINE LEARNING TO PREDICT HIGH-RISK CUSTOMERS

TABLE OF CONTENT

1.PROJECT GOAL

Predict whether a loan applicant will default or not.

Target column: TARGET

0 → No Default

1 → Default

2.DATASET OVERVIEW:

- Rows: 307511
- Features: 33
- Data Types:
- Financial information
- Demographic information
- Application/contract information

3.KEY COLUMNS EXPLAINED:

- AMT_INCOME_TOTAL → Total annual income of the client.
- AMT_CREDIT → Total credit amount of the loan.
- AMT_ANNUITY → Annuity amount to be paid every year.
- AMT_GOODS_PRICE → Price of the goods (if it's a goods loan).

- DAYS_BIRTH → Client’s age in days (negative values).
- DAYS_EMPLOYED → Days employed (negative values, special value 365243 = unemployed).
- EXT_SOURCE_1/2/3 → External credit scores from other sources.
- NAME_CONTRACT_TYPE → Type of loan (Cash/ Revolving).

- CODE_GENDER → Gender of the client.
- FLAG_OWN_CAR / FLAG_OWN_REALTY → Does the client own a car or property?
- CNT_CHILDREN → Number of children.
- CNT_FAM_MEMBERS → Family size.
- OCCUPATION_TYPE → Client occupation.
- NAME_HOUSING_TYPE → Housing situation (e.g. own, rent, parents).

DATA CLEANING:

- Checked for missing values → main columns with nulls: AMT_ANNUITY, EXT_SOURCE_2, CNT_FAM_MEMBERS, AMT_GOODS_PRICE, OCCUPATION_TYPE, NAME_TYPE_SUITE.

- Imputation strategy:

Numerical columns → filled with median (not normally distributed).

Categorical columns → filled with mode.

- Verified no null values remained.

5.TARGET IMBALANCE:

- Class distribution:
  0 → 282,686 clients
  1 → 24,825 clients

- Highly imbalanced dataset → used oversampling.

6.DATA PREPARATION:

- Converted specific columns to categorical type.
- Scaled numerical features using MinMaxScaler.
- Encoded categorical features with LabelEncoder.

7.BALANCING WITH SMOTE:

- Applied SMOTE oversampling:

Once before trying different models.

Again before final XGBoost training.

- This improved recall (detecting high-risk clients).

8.FEATURE SELECTION:

- Ran XGBoost on all features → got feature importance.
- Selected Top 10 most important features:

1. EXT_SOURCE_2
2. AMT_INCOME_TOTAL
3. DAYS_BIRTH
4. HOUR_APPR_PROCESS_START
5. NAME_CONTRACT_TYPE
6. FLAG_OWN_CAR
7. CNT_FAM_MEMBERS
8. AMT_GOODS_PRICE
9. REGION_POPULATION_RELATIVE
10. CNT_CHILDREN

9.MODEL TRAINING:

Models tested:

1. Decision Tree
2. Random Forest
3. XGBoost (final)

10.MODEL RESULTS:

Decision Tree

Train Acc: 0.857

Test Acc: 0.826

AUC: 0.887

Random Forest

Train Acc: 0.879

Test Acc: 0.846

AUC: 0.925

XGBoost (Best)

Train Acc: 0.9305

Test Acc: 0.9282

Precision: 0.973

Recall: 0.881

F1 Score: 0.925

AUC: 0.963

11.STREAMLIT APP:

Deployed the best model (XGBoost) using Streamlit.

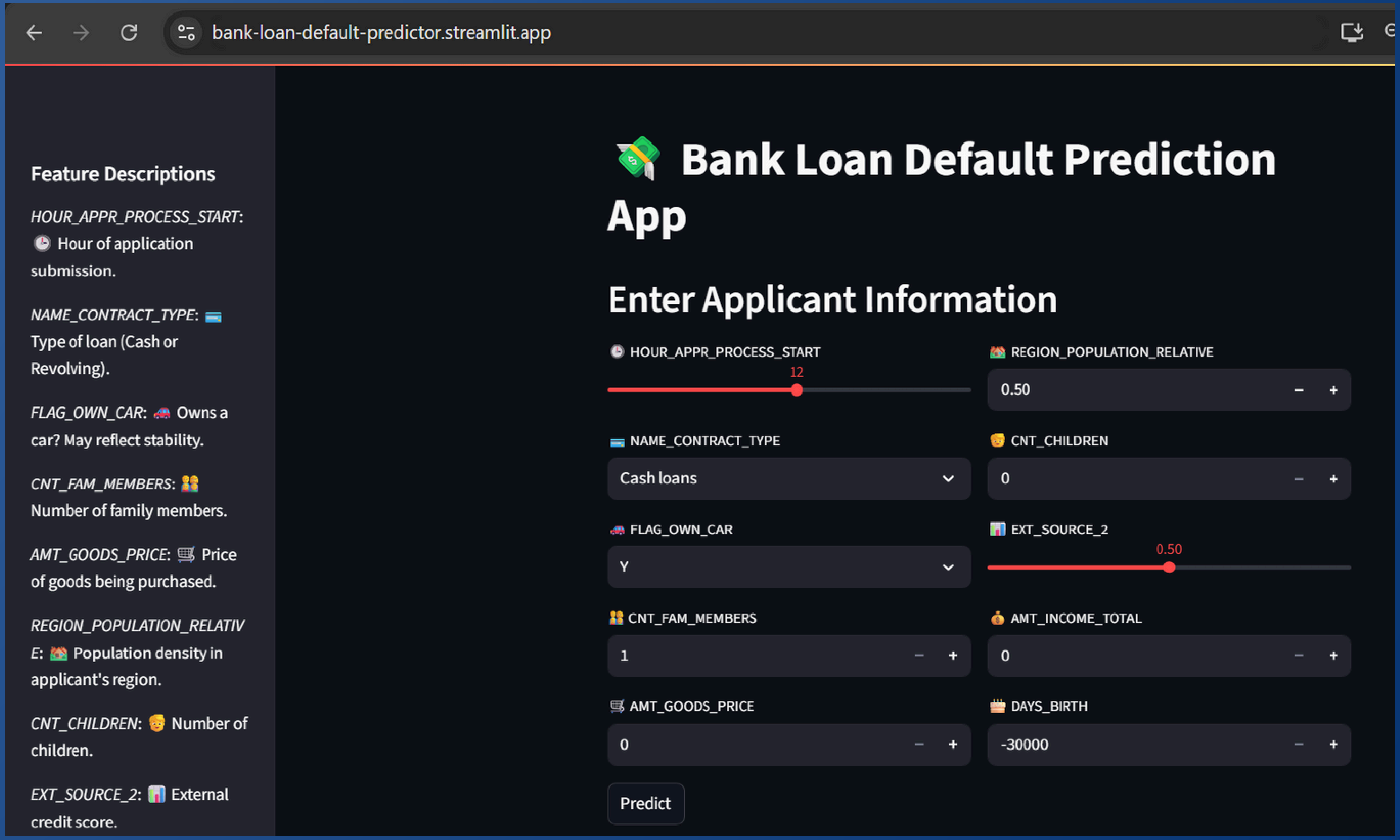
User inputs features → app predicts:

Probability of Low Risk

Probability of High Risk

Final prediction label.

website link :https://bank-loan-default-predictor.streamlit.app/



12.CONCLUSION:

- Data preprocessing (filling nulls, encoding, scaling) was essential.
- SMOTE balancing improved minority detection significantly.
- Feature selection simplified the model without major accuracy loss.
- XGBoost outperformed other models with 92.8% accuracy and 0.96 AUC.
- Fully deployed as an interactive web app for real-time predictions.

THANK YOU