Loan Default Model

- Yuke Gan, Ying Feng, Shihan Wang

Data

- The data comes from Kaggle:
  https://www.kaggle.com/datasets/hemanthsai7/loandefault/data
- Total 35 Columns:
    - Except ID and Loan Default target column, 33 Columns left
    - 9 Categorical Columns, 24 Numerical Columns

Data Preprocess

1. Load Data
2. Check Data
    a. Drop rows with nan values (No Nan Values)
    b. Drop feature with only one unique value (One column dropped)
3. Prepare Train and Test Dataset
    a. Get X, y dataset: X contains 32 features, y is loan default column (target value 0 & 1)
    b. Convert categorical columns to numbers by LabelEncoder()
    c. Check target value imbalance: 90% 0 and 10% 1
        i. Balance data by imblearn.over_sampling.SMOTE
    d. Split X_train, X_test, y_train, y_test with 80% train data, 20% test data
    e. Standardize the dataset by StandardScaler()

Pick Model

1. Train 5 different models by using Metaflow's foreach method
   a. RandomForestClassifier(random_state=42),
   b. XGBClassifier(random_state=42),
   c. AdaBoostClassifier(random_state=42),
   d. GradientBoostingClassifier(random_state=42),
   e. LogisticRegression(random_state=42)
2. Pick the model with highest **precision in target value 1**
   a. Using classification_report: RandomForestClassifier(random_state=42)

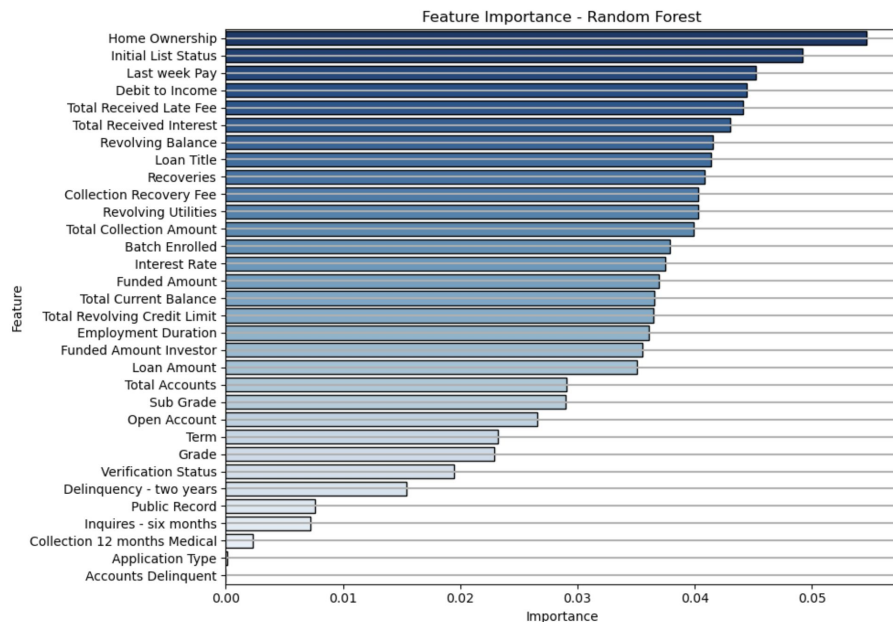|	Model	Precision on 1
0	Random Forest Classifier	0.893130
1	XGBClassifier	0.842205
2	AdaBoost	0.763595
3	Gradient Boosting	0.791810
4	Logistic Regression	0.668601

Improve Model

1. Check Feature Importance
   a. Model.feature_importances_ :
   b. Train RFC Again with updated features
      i. Original Precision: 0.893130

Feature Importance - Random Forest

```
{'0': {'precision': 0.8640967260736437,
  'recall': 0.8988158432013067,
  'f1-score': 0.8811144023697062,
  'support': 12245},
 '1': {'precision': 0.894571136827774,
  'recall': 0.858624632473048,
  'f1-score': 0.8762293715619269,
  'support': 12244},
 'accuracy': 0.8787210584343991,
 'macro avg': {'precision': 0.8793339314507089,
  'recall': 0.8787202378371773,
  'f1-score': 0.8786718869658166,
  'support': 24489},
 'weighted avg': {'precision': 0.879333309244601,
  'recall': 0.8787210584343991,
  'f1-score': 0.8786719867051038,
  'support': 24489}}
```

Improve Model

2. Search Best Performed Hyperparameters by RandomizedSearchCV:

    a. Pick model by target value 1's precision.

```python
param_dist = {
    'n_estimators': [10, 50, 100],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 4, 6],
    'min_samples_leaf': [1, 2, 3],
    }
precision_scorer = make_scorer(precision_score, pos_label=1)
rf_random = RandomizedSearchCV(estimator=rf,
                               param_distributions=param_dist,
                               n_iter=50, cv=2, verbose=2,
                               random_state=42, n_jobs=-1,
                               scoring=precision_scorer)
```

Improve Model

```
{'0': {'precision': 0.8551050469028607,
  'recall': 0.9007758268681094,
  'f1-score': 0.8773464842507158,
  'support': 12245},
 '1': {'precision': 0.8951682484900777,
  'recall': 0.8473538059457694,
  'f1-score': 0.8706050180414535,
  'support': 12244},
 'accuracy': 0.8740659071419821,
 'macro avg': {'precision': 0.8751366476964693,
  'recall': 0.8740648164069393,
  'f1-score': 0.8739757511460846,
  'support': 24489},
 'weighted avg': {'precision': 0.8751358297128524,
  'recall': 0.8740659071419821,
  'f1-score': 0.8739758887888266,
  'support': 24489}}
```

App Website

Using Flask to build the app website.

Predict

Predicted label: 0

Loan Default Classifier

Example Applicant's Characteristics:

Loan Amount: 32879
Funded Amount: 1014
Funded Amount Investor: 10989.93948
Term: 59
Batch Enrolled: BAT5629144
Interest Rate: 6.894023179
Grade: B
Sub Grade: A1
Home Ownership: MORTGAGE
Employment Duration: 94327.15399
Verification Status: Not Verified
Loan Title: Debt consolidation
Debt to Income: 34.8737539
Delinquency - two years: 0
Inquires - six months: 0
Open Account: 11
Public Record: 0
Revolving Balance: 2753
Revolving Utilities: 68.81130029
Total Accounts: 19
Initial List Status: w
Total Received Interest: 471.8253429
Total Received Late Fee: 0.092744198
Recoveries: 3.789364581
Collection Recovery Fee: 0.578289675
Collection 12 months Medical: 0
Application Type: INDIVIDUAL

Loan Amount: loan amount applied `Enter a numerical value`
Funded Amount: loan amount funded `Enter a numerical value`
Funded Amount Investor: loan amount approved by the investors `Enter a numerical value`
Term: term of loan (in months) `Enter a numerical value`
Batch Enrolled: batch numbers to representatives `BAT5629144`
Interest Rate: interest rate (%) on loan `Enter a numerical value`
Grade: grade by the bank `B`
Sub Grade: sub-grade by the bank `A1`
Home Ownership: Type of home ownership `MORTGAGE`
Employment Duration: Employment duration in hours `Enter a numerical value`
Verification Status: `Not Verified`
Loan Title: Purpose for loan `Debt consolidation`
Debt to Income: ratio of representative's total monthly debt repayment divided by self reported monthly income excluding mortgage `Enter a numerical value`
Delinquency - two years: number of 30+ days delinquency in past 2 years `Enter a numerical value`
Inquires - six months: total number of inquiries in last 6 months `Enter a numerical value`
Open Account: number of open credit line in representative's - credit line `Enter a numerical value`
Public Record: number of derogatory public records `Enter a numerical value`
Revolving Balance: total credit revolving balance `Enter a numerical value`
Revolving Utilities: amount of credit a representative is using - relative to revolving_balance `Enter a numerical value`
Total Accounts: total number of credit lines available in - representatives credit line `Enter a numerical value`
Initial List Status: unique listing status of the loan - - W(Waiting), F(Forwarded) `w`
Total Received Interest: total interest received till date `Enter a numerical value`
Total Received Late Fee: total late fee received till date `Enter a numerical value`
Recoveries: post charge off gross recovery `Enter a numerical value`
Collection Recovery Fee: post charge off collection fee `Enter a numerical value`
Collection 12 months Medical: total collections in last 12 months - excluding medical collections `Enter a numerical value`
Application Type: indicates when the representative is an individual or joint `INDIVIDUAL`
Last week Pay: indicates how long (in weeks) a representative has paid EMI after batch enrolled `Enter a numerical value`
Accounts Delinquent: number of accounts on which the representative is delinquent `Enter a numerical value`
Total Collection Amount: total collection amount ever owed `Enter a numerical value`
Total Current Balance: total current balance from all accounts `Enter a numerical value`
Total Revolving Credit Limit: total revolving credit limit `Enter a numerical value`

`Predict`