Predicting Loan Foreclosures

How Freddie Mac could reduce foreclosure losses by 80%

The problem

Company

Freddie Mac
We make home possible®

Freddie Mac operates in the U.S. secondary mortgage market. They don't lend directly to borrowers but buy loans that meet their standards from approved lenders.

Context

Freddie Mac provides Single Family Loan-Level Dataset in an effort to increase transparency and help investors build more accurate credit performance models.

I chose data from 1999 Q1 for my analysis.

Problem statement

About 1.6% of the loans in the 1999 loan dataset ended up in foreclosure. My goal was to understand how well modeling could predict foreclosure of loans. Potentially reducing loan loss for Freddie Mac

Challenges deep-dive

Challenge 1

Data Imbalance

Total records: 383,834

Normal Loans: 379,370

Foreclosed Loans: 4,464

RandomForestClassifier Accuracy > 98%

Challenge 2

Data Undersampling

Determine the right ratio for undersampling.

Records of normal loans

to

Records of foreclosed loans

Challenge 3

Modelling

RandomForestClassifier

KNeighborsClassifier

Support Vector Classifier

Hyperparameter Tuning

Cross Validation

Solution

Random Forest Classification method resulted in a recall score of 0.81

Performed modelling with

- Random Forest Classification
- K Neighbors Classifier
- Support Vector Classifier

Nested Cross Validation with

- GridSearchCV for Hyperparameter tuning
- Cross_val_score to measure the prediction performance of the estimator

Implementation

15 Features. Categories to numbers, fill missing values, remove columns with too many null values

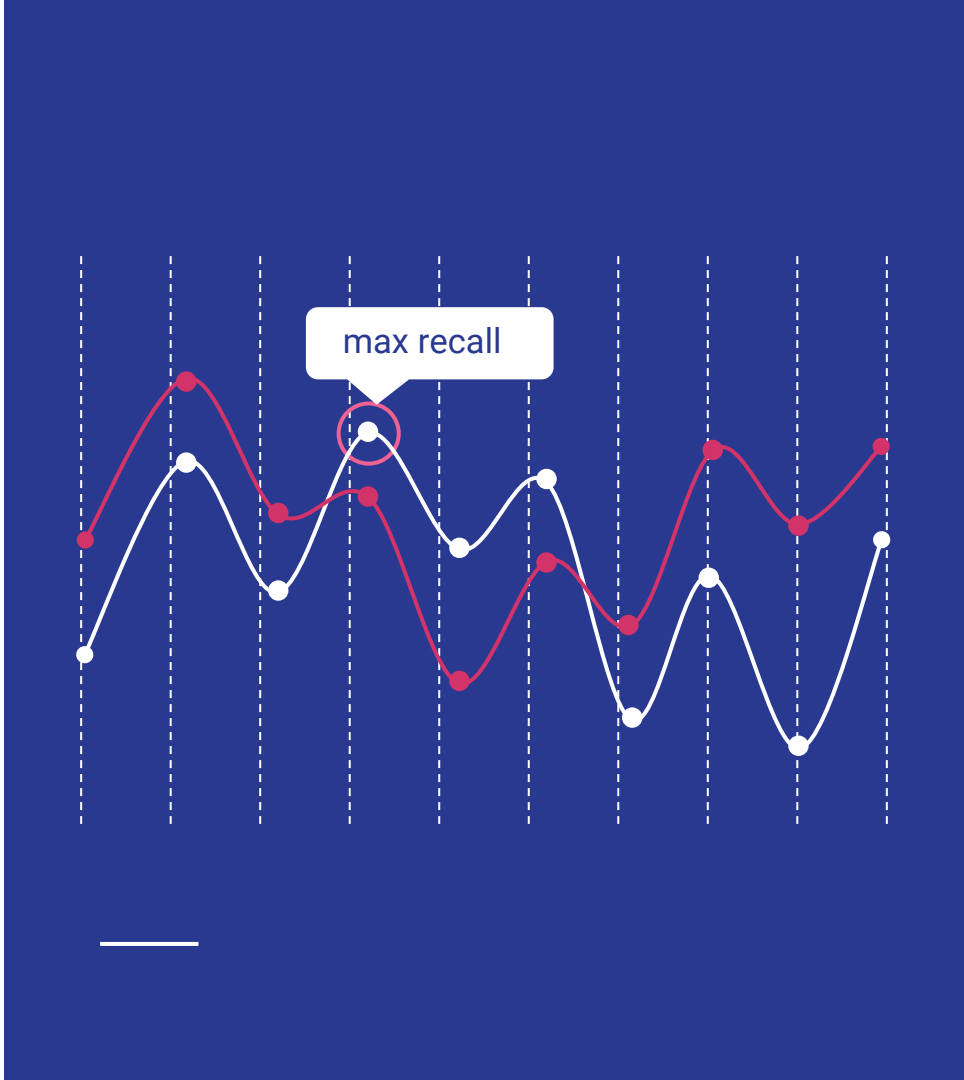
Use of nested cross validation combining GridSearchCV and cross_val_score

Experimented with 1:1, 2:1 and 3:1 ratios of normal to foreclosed

RandomForestClassifier
KNeighborsClassifier
SupportVectorClassifier

Impact

Potential loss reduction
$347,580,720



Implementation Details

EDA

Features:

- Credit Score
- First Time Buyer (category to bool)
- Number of Units (median to fill missing values)
- Occupancy Status (category to bool)
- Loan-to-value (4 missing values, removed rows)
- Debt-to-income (replaced missing values with median)
- Unpaid balance
- Interest rate
- Channel (category to int)
- Property state (category to int)
- Property type (category to int)
- Purpose (category to int)
- ...

Target:

- Foreclosure status (category to bool)

Undersampling

Normal Loans: 379,370 (98.84%)
Foreclosed Loans: 4,464 (1.16%)

RandomForestClassifier on original dataset:

- 98% accuracy
- Recall score of 0.56 !!

Determine Optimal Undersampling Ratio

Ratio	Algorithm	Recall score
1:1	Logistic Reg	0.75
1:1	Random Forest	0.78
2:1	Logistic Reg	0.43
2:1	Random Forest	0.59
3:1	Logistic Reg	0.23
3:1	Random Forest	0.46

Classification Algorithms

Nested Cross Validation for Hyperparameter Tuning and Prediction Performance

Random Forest Classifier:

- Hyperparameter Tuning:
  - N_estimators: 100
  - Max_depth: 10
  - Min_samples_leaf: 1
- Recall score: 0.81

KNeighbors Classifier:

- Hyperparameter Tuning:
  - N_neighbors: 5
  - leaf_size: 1
- Recall score: 0.73

Support Vector Classifier:

- Hyperparameters Tuning:
  - C: 6
- Recall score: 0.51