Project #2

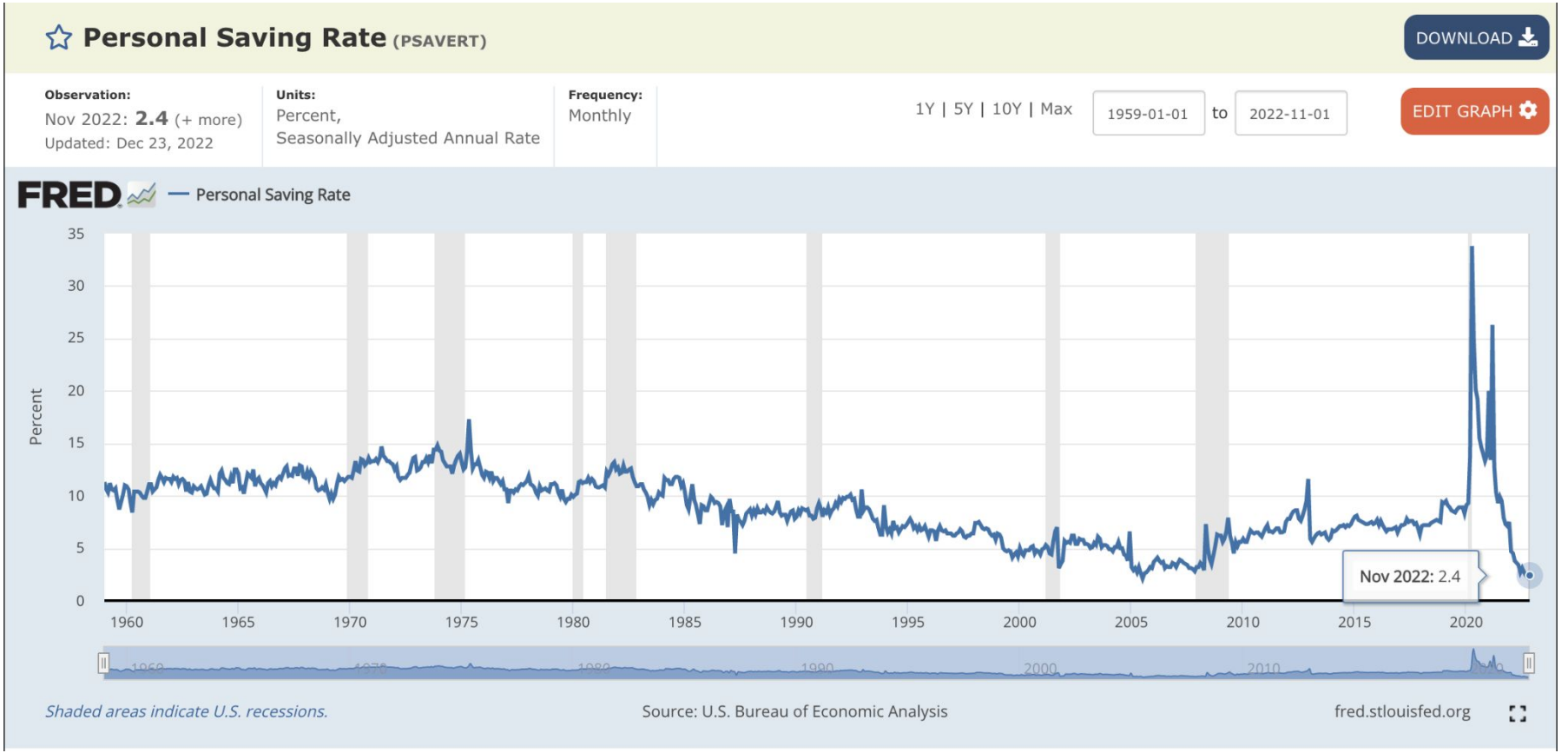
Cesar Sandiford

Loan Default Prediction

This project is about using machine learning to predict loan defaults with given data about customers with a Supervised model.

In times of uncertainty, FinTechs are in big risk when it comes to higher percentage of loan defaults and using machine learning to understand and identify the risk is very important to have in hand so they can make decisions to raise the credit standards in the future and minimize risk.

Inspired by:


Personal Saving Rate (PSAVERT)
DOWNLOAD
Observation:
Nov 2022: 2.4 (+ more)
Updated: Dec 23, 2022
Units:
Percent,
Seasonally Adjusted Annual Rate
Frequency:
Monthly
1Y | 5Y | 10Y | Max
1959-01-01
to
2022-11-01
EDIT GRAPH
FRED
Personal Saving Rate
Percent
35
30
25
20
15
10
5
0
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
Nov 2022: 2.4
Shaded areas indicate U.S. recessions.
Source: U.S. Bureau of Economic Analysis
fred.stlouisfed.org

Consumer Loans: Credit Cards and Other Revolving Plans, All Commercial Banks

(CCLACBW027SBOG)

DOWNLOAD

Observation:
2023-01-04: **940.0455** (+ more)
Updated: Jan 13, 2023

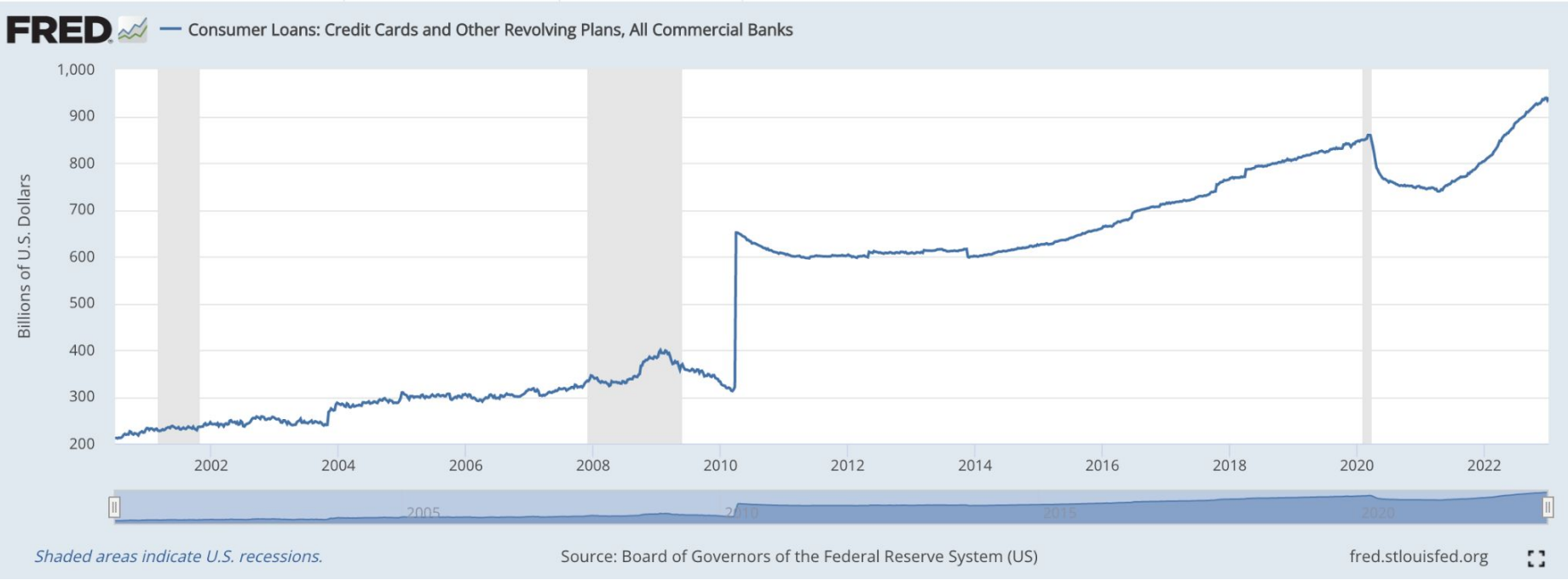
Units:
Billions of U.S. Dollars,
Seasonally Adjusted

Frequency:
Weekly,
Ending Wednesday

1Y | 5Y | 10Y | Max

2000-06-28 to 2023-01-04

EDIT GRAPH

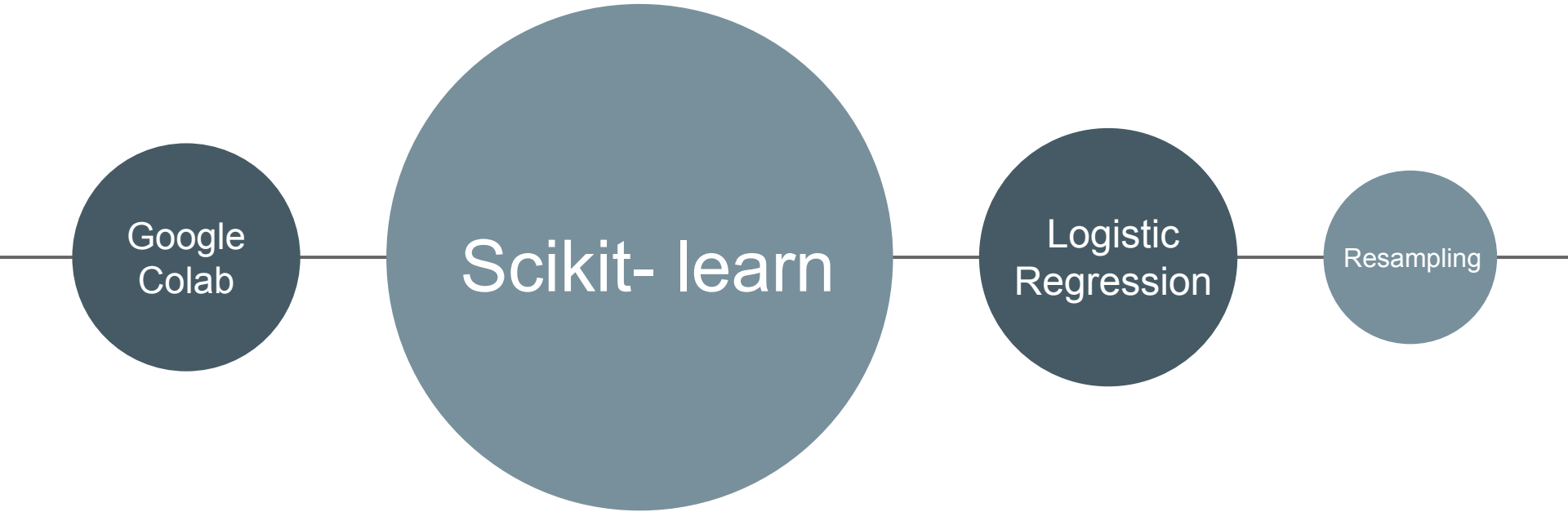


Shaded areas indicate U.S. recessions.

Source: Board of Governors of the Federal Reserve System (US)

fred.stlouisfed.org

Skills Used


Google Colab
Scikit- learn
Logistic Regression
Resampling

Walkthrough

Importing Libraries

Numpy

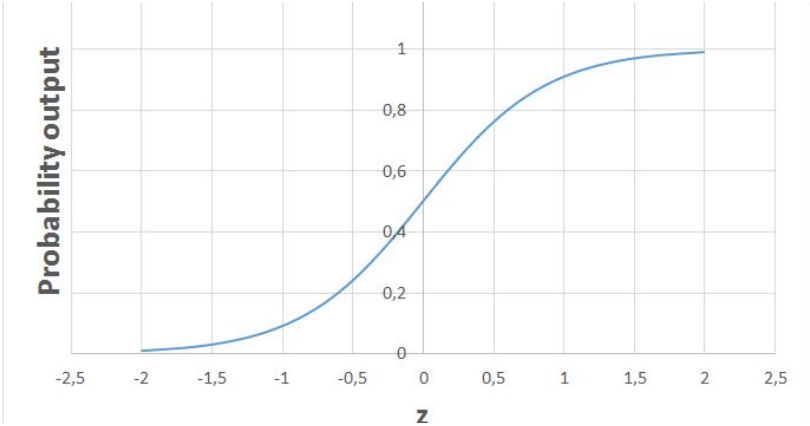
Pandas

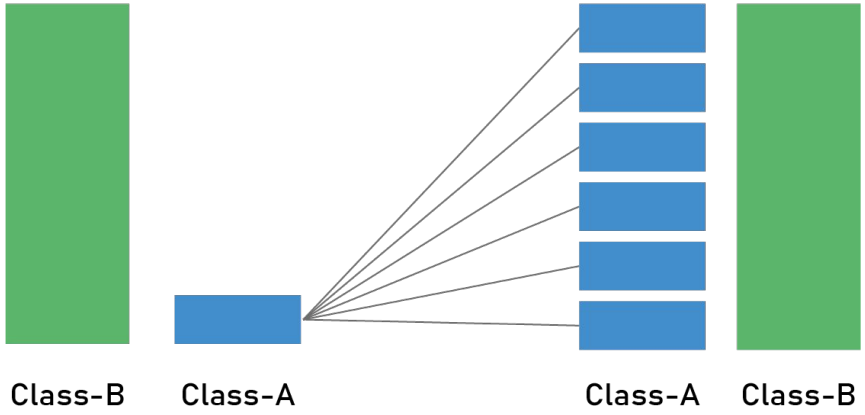
Pathlib-Path

Sklearn

Train-test-split

Classification report



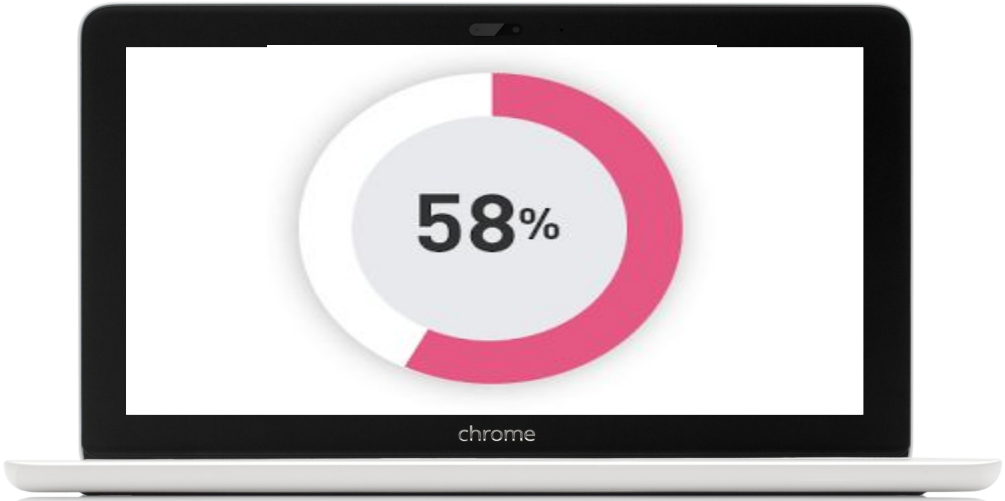


Logistic Regression

Logistic regression can be used to predict default events and model the influence of different variables on a consumer's credit- worthiness. In this paper we use a logistic regression model to predict the creditworthiness of bank customers using predictors related to their personal status and financial history.

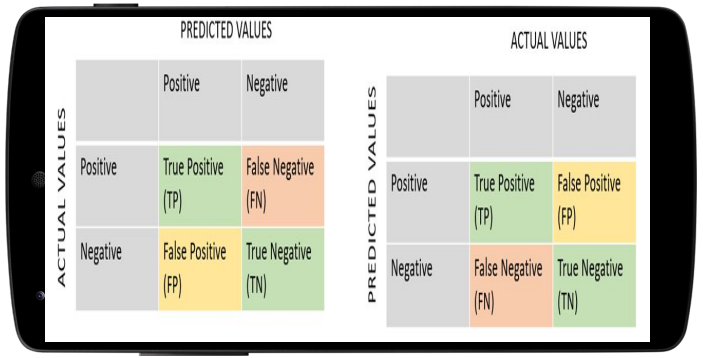
Resampling

By resampling our data, we have the chance to create better-performing machine learning models. And, the more resampling techniques that we have, the better our chances are of identifying one that works well for our current data. In this case, the oversampling was the one for this task.



Balance Accuracy Score

Balanced Accuracy is used in both binary and multi-class classification. Sensitivity and specificity, its use case is when dealing with imbalanced data, when one of the target classes appears a lot more than the other.



Confusion Matrix

It measures the quality of predictions from a classification model by looking at how many predictions are True and how many are False. A true positive is an outcome where the model correctly predicts the positive class. Similarly, a true negative is an outcome where the model correctly predicts the negative class. A false positive is an outcome where the model incorrectly predicts the positive class.

Conclusion

By having the Logistic Regression model applied to the dataset, the balanced accuracy score was 58% meaning it was not too accurate. Since the data was imbalanced by the defaults, I gave it a proper chance by oversampling it to increase the accuracy of the mode up to 84%. Huge improvement! F1 score average was 94 vs the resampled average was 90. In conclusion, Oversampling the model brings a better prediction for this situation. Also, predicting defaults can be difficult given that customers could have different lifestyle, emergencies, unexpected situation or money opportunities in life.

Credits

Kaggle:
https://www.kaggle.com/datasets/b934bbd9d19e1c321cf5f121b9b8f9bea4f1e770ea8d4cfdf445fb34ab80f42c?resource=download

https://www.kaggle.com/datasets/itssuru/loan-data

Module 12 Challenge

Google.com

Stack Overflow

Federal Reserve Economic Data: https://fred.stlouisfed.org/series/PSAVERT & https://fred.stlouisfed.org/series/CCLACBW027SBOG