# Supervised Machine Learning Project README

## Project Overview

This project aims to build supervised machine learning models to predict credit risk. The data provides data on loan size, interest rate, borrower income, debt to income ratio, number of accounts, derogatory marks, and loan status. Loan status is designated by ‘0’ (low risk) and ‘1’ (high risk). The goal was to predict loan risk.

The first model involved creating a dataframe from lending\_data.csv. The label column, ‘loan\_status’, was separated from the ‘features’ columns. The balance was a 75/25 split. Train\_test\_split method was used to create train and test sets for X and y. A regression was fit then passed through the ‘.predict ‘method. Results seen below.

The second model used RandomOverSampler to resample the data, accounting for any imbalance in the data. The data was then fit then passed through a LogisticalRegression. A fit with the resampled data was completed. A ‘.predict’ method was then used. Results below.

## Installation

To run this project, you need to have Python 3.7 or higher installed on your machine. Additionally, you'll need the following Python libraries:

* NumPy
* pandas
* pathlib
* sklearn.metrics - balanced\_accuracy\_score, confusion\_matrix, classification\_report
* sklearn.model\_selection - train\_test\_split
* sklearn.linear\_model - LogisticRegression
* imblearn.over\_sampling – RandomOverSampler

You can install these dependencies using pip:

pip install NumPy, pandas, scikit-learn, matplotlib

## Data

The dataset used for this project is stored in the `Resources’ folder. It consists of a CSV file named `lending\_data.’ The file contains the following columns, in order:

1. loan size: in whole US dollars (denomination is assumed in absence of any indicators)
2. interest rate: in percent
3. borrower income: in whole US dollars per year
4. debt to income ratio: monthly debt divided by gross monthly income (https://www.consumerfinance.gov/ask-cfpb/what-is-a-debt-to-income-ratio-en-1791/#:~:text=Your%20debt%2Dto%2Dincome%20ratio,will%20have%20different%20DTI%20limits). Ranges from 0 to 1, with 1 being high debt to income ratio. Lower is preferred.
5. number of accounts: number of credit and revolving credit accounts
6. derogatory marks: are negative reflections on your credit report indicating that one did not repay as agreed. Examples are late payments, charge-off of delinquent debt, or bankruptcy. This is not a complete list (<https://www.creditkarma.com/advice/i/what-does-derogatory-mean#:~:text=Derogatory%20marks%20are%20negative%2C%20long,reports%20as%20a%20derogatory%20mark)%>
7. loan status: 0 is labeled low-risk and 1 is labeled high-risk

## Data Pre-processing

Prior to model training, we perform the following data pre-processing steps:

- Handling missing values: not necessary; code was curated for class use

- Encoding categorical variables: All values were numerical

- Splitting the data: We split the dataset into training and testing sets to evaluate the model's performance.

## Model Training

Two supervised machine learning models were created. The first model was using the LogisticRegression method on the original data. The data was then resampled using RandomOverSampler to balance the data, then created a second LogisticRegression model.

## Results and Visualizations

Model 1: At 0.99% accuracy, the linear regression model demonstrates to be an accurate model when predicting high-risk loans; balanced accuracy was 0.95. The classification report looks excellent at 1.00 Precision and 0.99 Recall. In sum, the model appears to be an excellent predictor of high-risk loan categories.

Model 2: After balancing the data and completing another linear regression, all numbers were identical to the first model in the classification report and balanced average. This, too, appears to be an excellent model.

It is noted, however, that the two models produced different rates of false positive and false negative which could be used to make a final determination in which model to employ.

## Reproducibility

To reproduce the results of this project, simply run the `main.py` script provided in the repository. Make sure you have the required dependencies installed (see Installation section). The script will load the data, perform data pre-processing, train the model, and display the results.

## Acknowledgments

I would like to thank my instructors, BCS Support team and my tutor.

## Contact Information

For any questions or feedback regarding this project, feel free to contact Jason Stone at jasonstonemed@gmail.com.ith balanced data. The classification report looks excellent at 1.00 precision and 0.99 recall. In sum, the model appears to be an excellent predictor of identifying high-risk groups.

At 0.99 accuracy, this demonstrates to be an accurate model when predicting high-risk applicants; accuracy 0.95 with balanced data. The classification report looks excellent at 1.00 precision and 0.99 recall. In sum, the model appears to be an excellent predictor of identifying high-risk groups.

At 0.99 accuracy, this demonstrates to be an accurate model when predicting high-risk applicants; accuracy 0.95 with ba##