The Erdös Institute

Responsible Lenders

Data Science Bootcamp - Autumn 2023

Presented on December 1st, 2023

Team #: Craig Franze, Patrick Millican, Chao Sun, Alex Melendez, André Martins

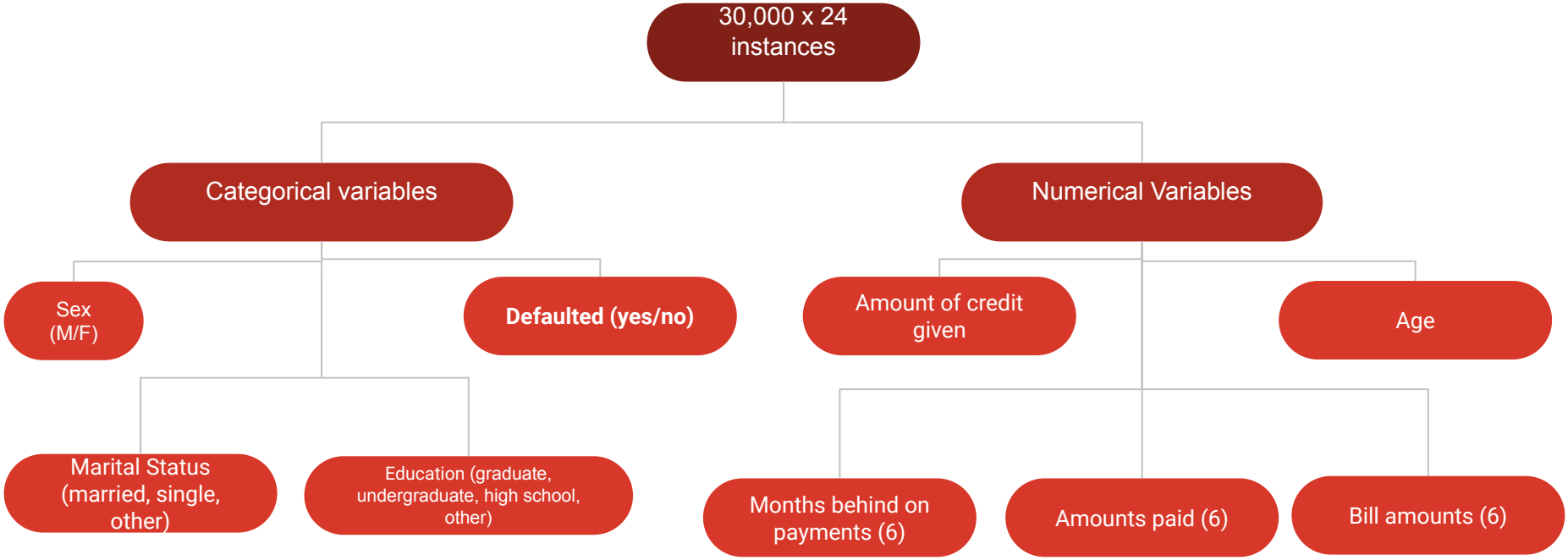
Github: https://github.com/Responsible-Lenders/credit-default

Project Overview

- Accurately predicting the probability of defaulting on credit cards is crucial for lenders to estimate risk and proportion interest rates accordingly—failure to do so may result in significant losses
- Aggregated personal data collected from borrowers can be correlated with the likelihood of default which can be modelled with trained classification algorithms
- For this project, we are using a dataset from UC Irvine Data Repository containing 30,000 instances of credit card default/lack of default in Taiwan over a 6-month period in 2005
- 23 variables associated with each instance are used to predict the likelihood of default

Dataset: Yeh,I-Cheng. (2016). default of credit card clients. UCI Machine Learning Repository. https://doi.org/10.24432/C55S3H

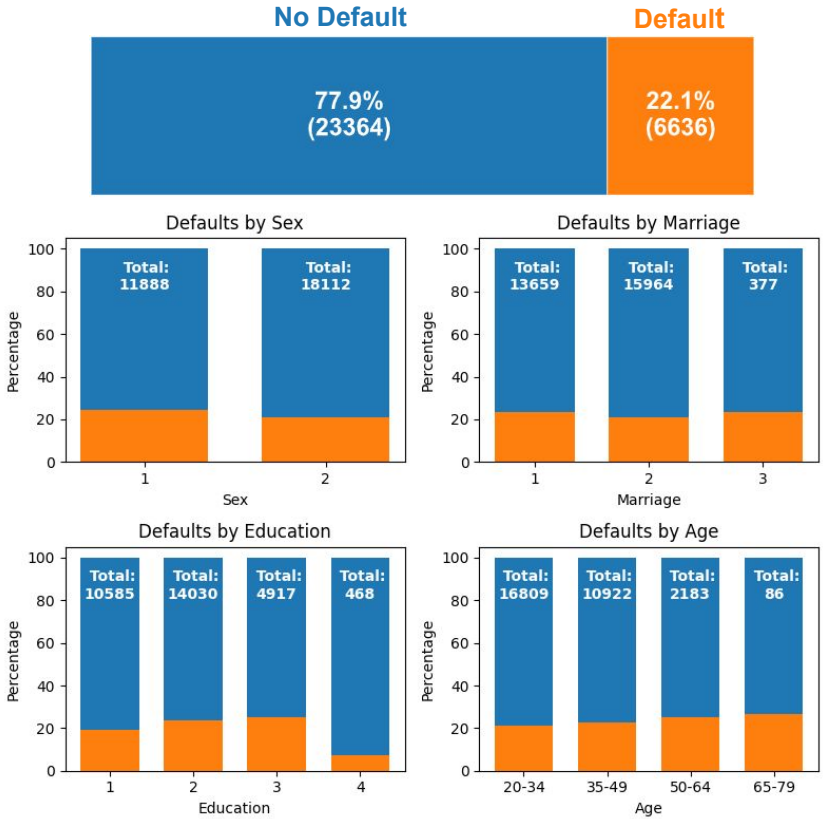
Dataset Features



Dataset: Yeh,I-Cheng. (2016). default of credit card clients. UCI Machine Learning Repository. https://doi.org/10.24432/C55S3H

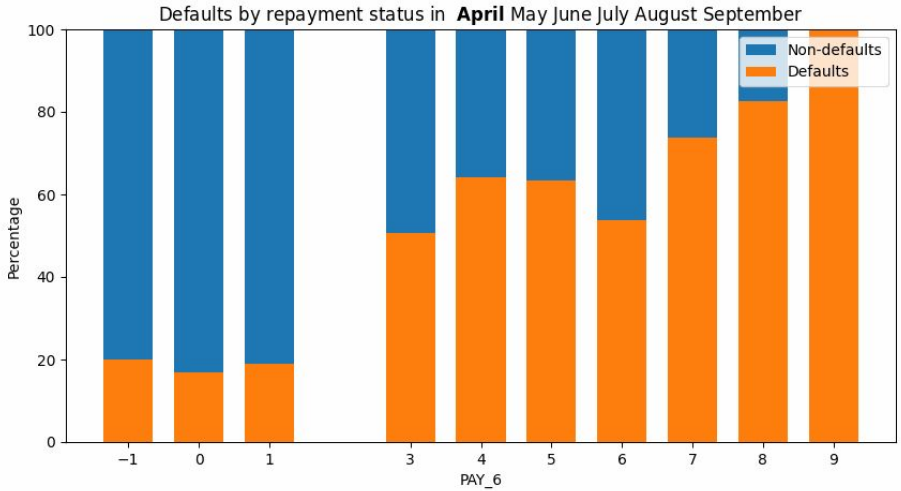
Exploratory Data Analysis

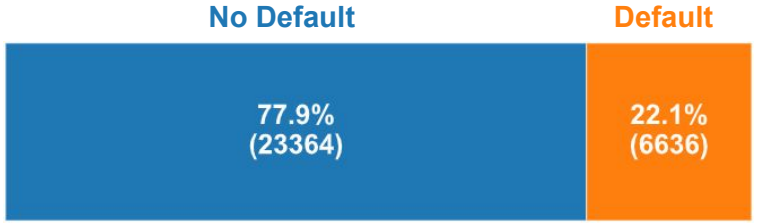
- Variables like Sex, Marriage, Education and Age barely influence Default probability
- Some categories have really low representation
- Class imbalance in default rate required data stratification



Exploratory Data Analysis

- Months behind on payments show highest correlation with default rate
- Especially for the last month, Pay_1
- Credit given, education, and age respectively have next highest correlations with default





Model Selection: XGBoost

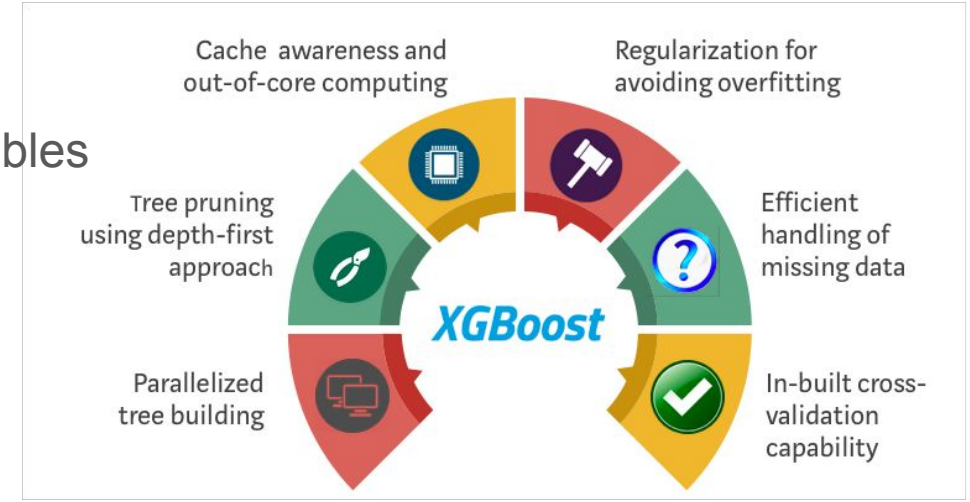
XGBoost

Advantages

- Efficient tree-based ML algorithm
- Scaling the data is not necessary
- Native support for categorical variables
- Hyperparameters can address class-imbalance

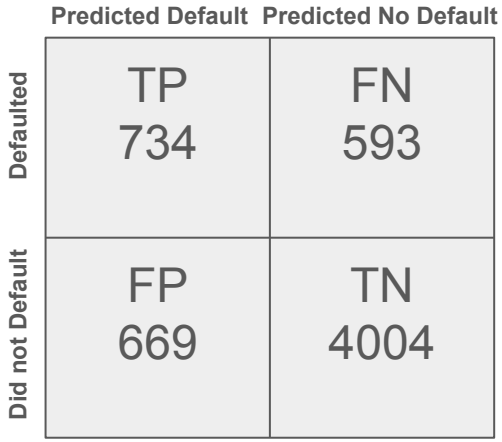
Disadvantages

- Algorithm is not easily explainable
- May be prone to overfitting
- Sensitive to outliers



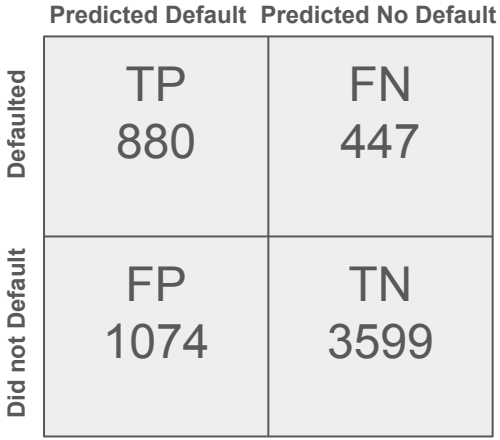
Results of XGBoost

- XGBoost (Recall driven)
  66% Recall, 45% Precision
  74.65% Accuracy
- XGBoost (Precision driven)
  38% Recall, 68% Precision
  82.23% Accuracy
- XGBoost (F1 driven)
  55% Recall, 52% Precision
  78.97% Accuracy



Balanced
Recall/Precision

Best Recall



Model Selection: Logistic Regression

Pros:

- Simple algorithm with interpretable output

Cons:

- Simplicity may cause algorithm to be outperformed by more complex algorithms

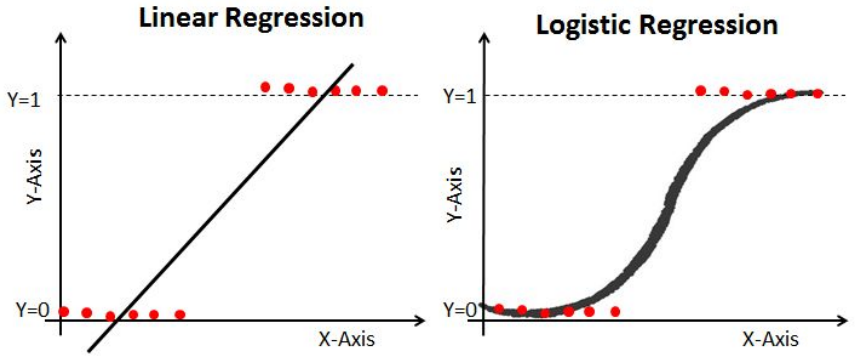
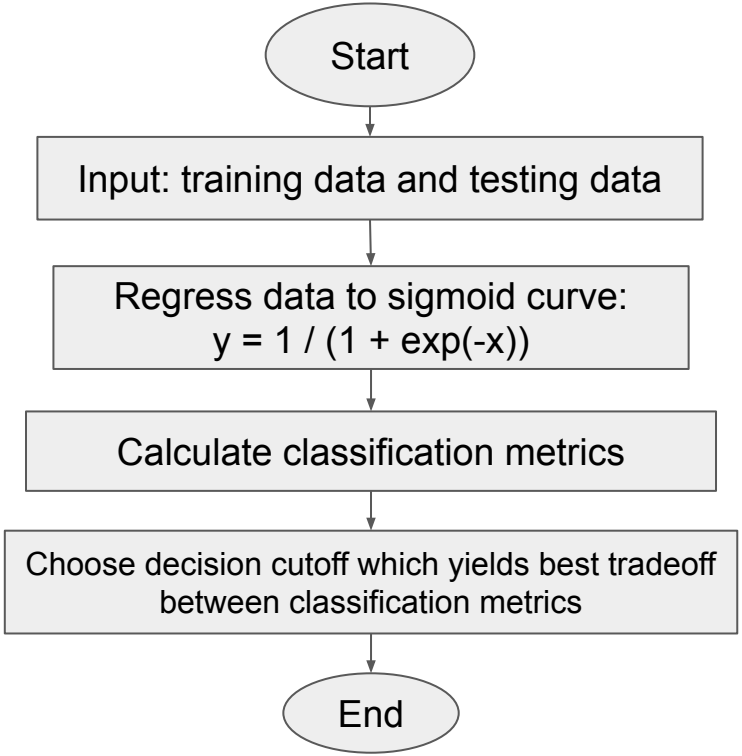


Image Source: www.datacamp.com

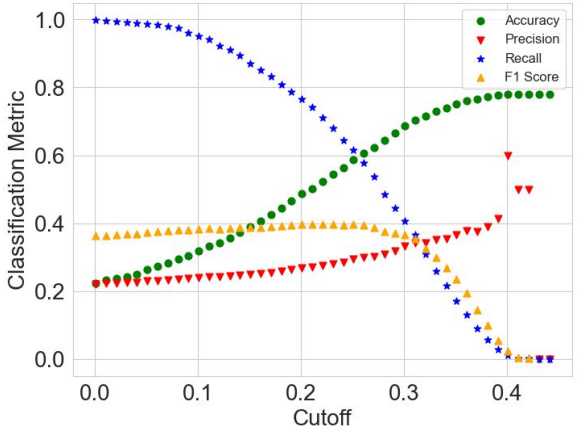


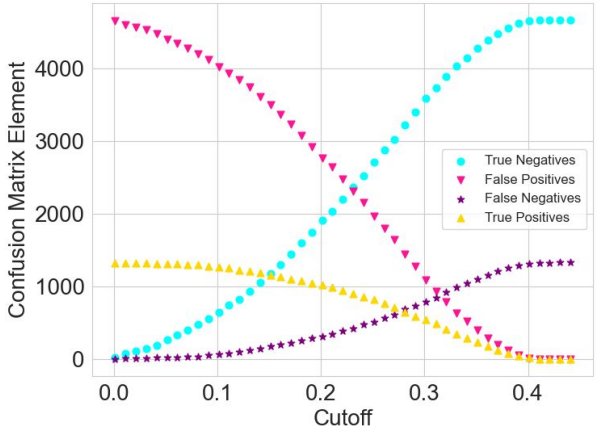
Results of Logistic Regression

`sklearn.linear_model`.LogisticRegression

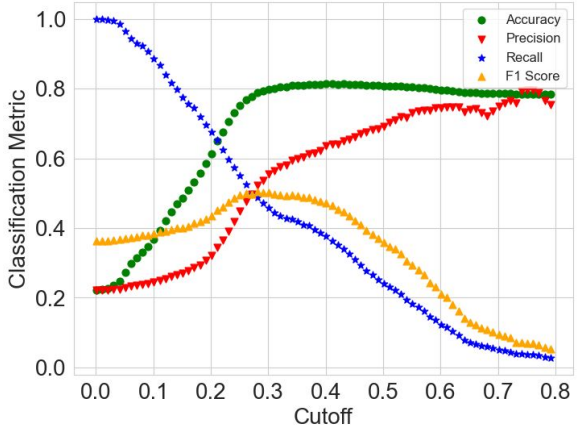
- Two logistic regression algorithms were compared: “L-BFGS” vs “liblinear” as functions of decision cutoff
- Overall liblinear performed better, with F1 Score and Accuracy maximized when cutoff values are within 0.25-0.4
- Constitutes 3% increase in Accuracy and 50% increase in F1 Score compared to predicting non-default in every instance.

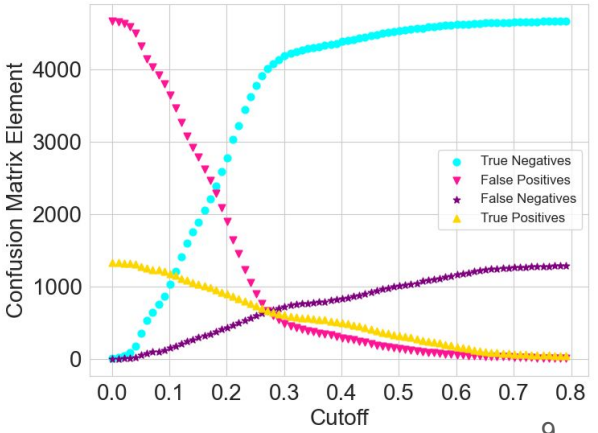
L-BFGS





liblinear





Project Results Summary

- More important
  - Confirms naive assumption that recent payment history best predicts default likelihood
  - Choice of model: XGBoost better balances recall and precision than logistic regression does, but both show promise
  - Education and marital status
- Less important
  - Rebalancing data
  - Age
  - Loan size
- Trade-offs
  - Precision and recall can be maximized separately
- Needs
  - Other data on borrowers for a stronger model

Real-World Impact and Future Applications

Our work makes bank lending more responsible for both lenders and borrowers by enabling informed pre-loan assessment of default likelihood, but there is still room for improvement in the quantity, timeliness, and breadth of data for training the model

Dataset: Yeh,I-Cheng. (2016). default of credit card clients. UCI Machine Learning Repository. https://doi.org/10.24432/C55S3H

Chao Sun
treychaosun@gmail.com

Craig Franze
craig.s.franze@gmail.com

Alex Melendez
alexleemelendez@gmail.com

Patrick Millican
pjmillican7@gmail.com

Andre Martins
andre02@ucsb.edu