LOAN DEFAULT PREDICTION .

Prinslou Tare
prinslou@stanford.edu

Robert Kinini
rkinini@stanford.edu

Ken Nturibi
knturibi@stanford.edu

Motivation

Following the Financial Crisis of 2008, credit institutions faced a number of loan defaults. This Jeopardized the profitability of banks due to number of nonperforming loans. Following these developments, there is a greater need for models that not only detect the risks at higher accuracy, but also deliver the detection at an early stage..

Problem Definition

The goal is to user a borrower's finance history to predict whether an individual is likely to default on a loan or not. The input is given as financial history and the output is either a 0 to mean 'paid off' or 1 to mean 'defaulted'

Features

The original dataset had 132 features and 819,501 observations. We removed irrelevant, scarce and protected features and retained 24 best features. E.g mortgage account, annual_income, empl_length e.t.c

Feature Encoding

We encoded the categorical features using feature hashing and one hot encoding.

For feature hashing, we hashed the zip codes in a smaller set of finite integer values and fed these values to our model.

We used the one hot encoding scheme to transform each attribute into m binary features where the label corresponding to the attribute is encoded as 1 and the rest are zeros.

Models/Approaches

Training and Test Size
Training set: ~ 655600
Test set: ~ 163901

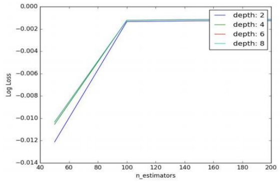
Logistic Regression
For logistic regression we setted for a linear solver and balanced class Weights

Random Forest
The random forest model used 100 estimators since that's what helped us achieve the best accuracy.

Light GBM
This implementation makes use of binary log loss and very low learning Rate of 0.001

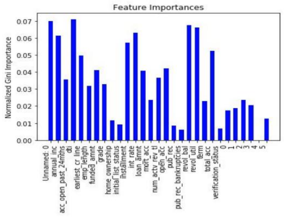
Multi Layer Perceptron
For MLP we used ReLU activation and adam optimization

XGB Boost
For this implementation we decided to use early stopping so as not to overfit the data



XGB Boost Loss Function

Feature Importance
We also analyzed feature importance and got the following results to the right



Results/Analysis

Model	Logistic Regression	MLP	Random Forest	Xgb Boost	Light GBM
Precision	0.79	0.78	0.78	0.78	0.78
Recall	0.67	0.82	0.82	0.82	0.82
Accuracy	0.67	0.81	0.81	0.82	0.82

Challenges

- There were alot of empty or sparse columns and rows which needed intensive prepossessing and feature engineering.
-There were a lot of anonymous features , e.g zip code, which limited the full potential of feature encoding and evaluation
-There is a huge disparity in the number of defaulted and number of paid off loans in out dataset.

Future

- We plan to under-sample negative examples so that we end up with an equal number of positive and negative examples.
- Since under-sampling may result in fewer training examples overall, another approach that can be taken is collecting more data for the positive examples so that we have an almost equal representation of positive and negative classes in our dataset.