Credit Default Predictor

CS 412 - Machine Learning - Final Project

University of Illinois at Chicago

Giuseppe Cerruto *(gcerru2@uic.edu)*
Edoardo Stoppa *(estopp2@uic.edu)*
Cosimo Sguanci *(csguan2@uic.edu)*

UIC

Introduction

- **Objective: predict if a customer will default on their bank payments using four classification ML techniques.**
- **Dataset:**
  - Number of instances: 30 000
  - Number of attributes: 23 (personal information, history of past payments, amount of given credit, previous bill statements, …)
- **Target**: binary value (1 if customer will default, 0 if not)

Dataset preprocessing

- **Originally 77.88% of the data belonged to class 0**

- Therefore, we balanced the data in order to have an even distribution between the two classes
- From 30 000 unbalanced instances, **we extracted 13 200 balanced instances**
  - 10 000 were used for training
  - 3 200 were used as test set

Applied Techniques

- **Logistic Regression & Support Vector Machine (Classifier)**
  - Selected because they are standard classification techniques
- **Naive Bayes**
  - Selected as a representative of generative classification techniques
- **Random Forest**
  - Selected because we wanted to explore a new technique which is not explained in the course

Accuracy

	Accuracy on training set (default)	Accuracy on test set (default)	Accuracy on training set (after CV)	Accuracy on test set (after CV)
Logistic Regression	61.13%	61.97%	67.56%	67.81%
Support Vector Machine	61.09%	60.78%	**74.17%**	**70.09%**
Gaussian Naive Bayes	54.32%	53.34%	55.40%	55.09%
Bernoulli Naive Bayes	67.89%	67.19%	68.39%	68.19%
Random Forest	**100%**	**69.87%**	71.85%	70.91%

Training and Inference speed

	Training time (ms)	Inference time (ms)
Logistic Regression	121.037	0.967
Support Vector Machine	2776.641	1377.313
Gaussian Naive Bayes	4.072	1.011
Bernoulli Naive Bayes	3.152	1.017
Random Forest	4147.934	107.023

Bias Test

	Accuracy on test set (after CV)	Accuracy after dropping “SEX” feature	Accuracy after dropping “PAY_*” features
Logistic Regression	67.81%	67.31%	61.94%
Support Vector Machine	70.09%	69.84%	61.94%
Gaussian Naive Bayes	55.09%	55.09%	55.09%
Bernoulli Naive Bayes	68.19%	67.87%	58.41%
Random Forest	70.91%	70.78%	65.38%

Thanks for your attention!

Giuseppe Cerruto *(gcerru2@uic.edu)*
Edoardo Stoppa *(estopp2@uic.edu)*
Cosimo Sguanci *(csguan2@uic.edu)*

Github Repository: **CreditDefaultPredictor (github.com)**