risks

MDPI

Article

Peer-To-Peer Lending: Classification in the Loan Application Process

Xinyuan Wei [1,2,*], Jun-ya Gotoh [3] and Stan Uryasev [2,*]

[1] School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China

[2] Risk Management and Financial Engineering Lab, Department of Industrial and Systems Engineering, University of Florida, 303 Weil Hall, Gainesville, FL 32611, USA

[3] Department of Industrial and Systems Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan; jgoto@indsys.chuo-u.ac.jp

* Correspondence: xinyuanwei@mail.dlut.edu.cn (X.W.); uryasev@ufl.edu (S.U.); Tel.: +86-13614112900 (X.W.); Tel.: +1-352-294-7723 (S.U.)

Received: 20 October 2018; Accepted: 5 November 2018; Published: 9 November 2018

Abstract: This paper studies the peer-to-peer lending and loan application processing of LendingClub. We tried to reproduce the existing loan application processing algorithm and find features used in this process. Loan application processing is considered a binary classification problem. We used the area under the ROC curve (AUC) for evaluation of algorithms. Features were transformed with splines for improving the performance of algorithms. We considered three classification algorithms: logistic regression, buffered AUC (bAUC) maximization, and AUC maximization.With only three features, Debt-to-Income Ratio, Employment Length, and Risk Score, we obtained an AUC close to 1. We have done both in-sample and out-of-sample evaluations. The codes for cross-validation and solving problems in a Portfolio Safeguard (PSG) format are in the Appendix. The calculation results with the data and codes are posted on the website and are available for downloading.

Keywords: peer-to-peer lending; loan application process; AUC maximization; bAUC maximization; spline approximation

1. Introduction

Peer-to-peer lending, also known as person-to-person lending, social lending, or P2P lending, commonly abbreviated as P2PL, is usually the online practice of individuals lending money to other individuals without going through a traditional financial intermediary. Basically, it involves people with extra money (investors), people who need money (borrowers), and a platform (website) that facilitates P2PL (see LendingClub (2006)).

After a borrower receives the money, monthly payments are made back to lenders through the platform. Some fraction of these payments is taken by the P2PL platform for providing the service. The most important task of a platform is to distinguish loan applicants who will pay money back from those who will default. This activity is called loan application processing.

This paper studies the loan application process of a P2PL company's platform and has two main objectives:

Objective 1. The main objective of this paper is to help investors using P2PL companies to understand the loan approval/classification process. With a practical example, we study classification algorithms for selecting loans. The case study was done for the LendingClub platform, which is a leading global P2P lending company. LendingClub was selected because of publicly available historical loan data. The comparison of the loan selection process of LendingClub with other companies is beyond the scope of this paper. While investors use ratings set by P2P lending companies for investment decisions, those companies do not provide detailed information on how ratings are set, and investors do not have appropriate information. We want to reduce this gap and explain with a case study that loan selection decisions are based on only several simple features (factors).

Objective 2. We want to demonstrate new classification techniques using spline transformation of features in combination with logistic regression, maximization of area under the ROC curve, and maximization of Buffered AUC.

We considered the loan applications process as a binary classification problem and used the AUC to evaluate algorithms. For conducting a case study, we used the open data of LendingClub and, in particular, features that were available for both approved and declined loans. Features were transformed with splines to improve classification algorithm performance. We conducted logistic regression with original and transformed features. Then, we maximized bAUC and AUC with transformed features. We applied the Portfolio Safeguard (PSG)[1] package for spline fitting and optimization. With only three features, Debt-to-Income Ratio, Employment Length, and Risk Score, we got an AUC close to 1.

Below are several of the main findings of the paper. A lot of information is collected and available for the evaluation of loans. However, most of this information is not used, and decisions are frequently based only on few key features (in the considered case study, only three features: Debt-to-Income Ratio, Employment Length, and Risk Score). We also found that popular simple technologies, such as logistic regression, are used for classification decisions. Some additional improvements can be obtained by using advanced algorithms, such as spline transformation of features and direct maximization of AUC and bAUC. These innovations can bring, in some cases, additional improvements, compared to basic simple technologies. Potential benefit for companies from this research is that expensive departments selecting loans can be substituted by a relatively cheap commonly used technologies (e.g., logistic regression with some additional innovations such as spline transformation of features). For P2PL investors this is also an important finding because it provides information about classification decisions used in practice. One more insight is that a standard PC can handle quite large datasets (with hundreds of thousands of observations) and advanced numerical capabilities (such as parallel processing) are not needed for loan selection.

The remaining part of the paper is structured as follows. Section 2 gives background information about P2PL. Section 3 introduces the loan application process and performance metric. This section also provides mathematical problem statements. Section 4 presents a case study. Section 5 concludes the paper.

2. Background

2.1. Peer-To-Peer Lending Companies

A P2PL company plays the same role in the P2PL market as the stock exchange in a stock market, and often has an online platform. ZOPA[2], the first company that offered P2P loans in the world, was founded in Britain in 2005. The name ZOPA, stands for "zone of possible agreement," a negotiating term identifying the bounds within which agreement can be reached between two parties (see Lai and Turban (2008)).

The first P2PL company in the United States, PROSPER Marketplace[3], a San Francisco, California-based company, was founded in 2005. PROSPER began operations in February 2006 and was the only P2PL company in the U.S. until LendingClub was founded in May 2007 in San Francisco, California. PROSPER was temporarily shut down in 2008 because of Securities and Exchange Commission (SEC) scrutiny. The majority of PROSPER investors got negative returns, mainly because of a poor loan evaluation model. PROSPER issued loans to anyone who had an interest in getting a loan.

[1] Portfolio Safeguard (PSG), http://www.aorda.com.
[2] ZOPA, http://www.zopa.com/.
[3] PROSPER Marketplace, https://www.prosper.com/.

Smith (1999) stated that the SEC issued its formal cease-and-desist letter, explaining that PROSPER is as a seller of securities and should be regulated by the SEC.

LendingClub was launched at first as a Facebook application. Within a couple of months, it emerged as a standalone website[4]. LendingClub was the first P2PL company who registered its offerings as securities with the SEC, and offered loan trading on a secondary market (run through a company called Foliofn[5]). Currently, it is the world's largest P2PL platform.

2.2. How Does It Work?

When someone needs a loan, they submit an application to a P2PL platform and become a potential borrower. The application includes information about the loan and the borrower, such as the amount requested, employment status, and social security number.

The platform accesses the status of a potential borrower using the Fair Isaac Credit Organization (FICO) score, debt-to-income ratio (DTI), home ownership, employment status, and other information. The platform decides whether to approve or decline a loan and sets an interest rate based on this information. The decision process is called loan application processing.

Once a loan is approved, potential lenders have 14 days to review the loan information and make an investment decision. The loan is issued if it receives enough funding within this period. The borrower receives money and makes monthly payments until they off off the loan. According to Lending Academy (2010), the lenders collect these payments minus a fee to the platform.

2.3. Loan Application Processing

The loan application process is a crucial procedure for a platform. This paper considers the loan application processing for the LendingClub. The company provides public access to approved/declined loan data and statistics. Initially, the loan applications go through a credit screening procedure. The applications passing the initial screening are evaluated by LendingClub's proprietary scoring models. The scoring model provides each applicant with a score, which is combined with the FICO score and other features. LendingClub considers about 180 features to decide whether to approve or decline a loan. For more details, we refer to LendingClub (2006).

2.4. Related Works on Peer-To-Peer Lending

As a novel financial model, P2PL has been extensively studied in the past two decades. Hulme and Wright (2006) focused on online social lending and provided an in-depth exploration of social lending from multiple perspectives; while Wang et al. (2009) provided an overview of the concept of P2PL, and discussed different P2PL marketplace models. Berger and Gleisner (2009) analyzed the role of a P2PL platform and found that market participants act as financial intermediaries and significantly improve borrowers' credit conditions by reducing information asymmetries. Lin (2009) investigated the role of "hard credit information" and "soft credit information" in the P2PL market. He found that loan applications with lower credit scores are less likely to be funded and more likely to default. Further, Iyer et al. (2009) found that a third of the variation in creditworthiness captured by the borrower's credit score can be inferred from available information. Puro et al. (2010) introduced a borrower decision-aid system that helps to formalize the decision-making process. Collier and Hampshire (2010) found that both loan amount and debt to income ratio of a borrower have influence on the final interest rate of a loan. Wu and Xu (2011) proposed a decision-support system providing individual risk assessment, eligible lender search, lending combination, and loan recommendation. Lin et al. (2013) studied friendship networks and information asymmetry in online P2PL and concluded that friendships increase the probability of successful funding, decrease interest

[4] LendingClub, https://www.lendingclub.com/.
[5] Foliofn, https://www.folioinvesting.com/folioinvesting/home/.

rates, and are associated with a lower ex post default rates. Chen et al. (2014) empirically tested data from PPDai and showed that both trust in borrowers and in intermediaries are significant factors influencing lenders' lending intention. Tsai et al. (2014) employed four machine-learning algorithms to classify and optimize peer lending risk, and found out that logistic regression outperformed LibSVM, Naïve Bayes, and random forest. Emekter et al. (2015) stated that higher interest rates charged on the high-risk borrowers are not enough to compensate for higher probability of default and claimed that, in order to sustain the business, LendingClub must attract borrowers with a high FICO score and high-income. Ma et al. (2017) studied different pricing mechanisms in peer-to-peer lending market, under the consideration of lenders' risk appetite. They have included the borrower pricing mechanism (BPM), the auction pricing mechanism (APM), which is Prosper's pricing mechanism before 2010, and the platform pricing mechanism (PPM), which Prosper used after a regime change. They claimed that, as long as the loan is profitable, the BPM and PPM are incentive-compatible mechanisms, while APM in not. Taking into consideration soft factors, Mi et al. (2018) established a model called SoFa to help investors to estimate default risks. Jiang et al. (2018) studied loan defaults by combining soft information extracted from descriptive text in online P2PL. Ding et al. (2018) investigated the lending transactions on RRDAI[6], and found a reputation mechanism that borrowers with better historical performance have a higher probability to obtain loans and at lower cost. Yu et al. (2018) studied the underlying neutral basis of herding behavior in online P2P lending at decision-making stage and feedback stage. By introducing event-related potentials (ERPs), they stated herding decision in P2PL is an evaluation of potential risk and it is effective for P2P platforms to optimize disclosure interfaces.

3. Methods and Performance Metric

The loan application process is a binary classification procedure that classifies a given set of loan applications into two classes (approved and declined loans).

Let $\{(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_m, y_m)\}$ be a set of m labeled loans, where $\boldsymbol{x}_i \in \mathbb{R}^k$ is the vector of features for a loan application i, and $y_i \in \{0, 1\}$ is the binary label of the loan application, where y equals 1 for approved and 0 for declined loans. Note that the dimension k of $\boldsymbol{x}_i$ is the number of features (also called credit attributes) provided by an applicant and a credit bureau. The loan application process can then be modeled as an estimation of a function $f : \mathbb{R}^k \to \{0, 1\}$, which is called a binary classifier, by using the existing labeled loan data.

3.1. Logistic Regression—A Benchmark

Logistic regression is a popular binary classifier suggested by Cox in 1958, see Freedman (2009). It assumes that given a feature vector $\boldsymbol{x} \in \mathbb{R}^k$, the probability that label $y = 1$ is given by

$$\Pr\{y = 1|\boldsymbol{x}\} = \frac{1}{1 + \exp(-S(\boldsymbol{x}))}, \tag{1}$$

where S is a function on $\mathbb{R}^k$, called a score function. A simple example of S is a linear function, $S(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$, where $\boldsymbol{w} \in \mathbb{R}^k$ and $b \in \mathbb{R}$ are parameters. These parameters can be estimated by the maximum likelihood method (see, for instance, Habermann 1979 and Hosmer et al. 2013). The case study in Section 4 uses logistic regression as a popular benchmark method.

3.2. Our Approach

In addition to a simple logistic regression we considered the following two-step procedure:

1. Each feature is transformed for finding the nonlinear dependence of the likelihood of loan approval using a feature-wise spline regression.

[6] www.renrendai.com.

2. Nonlinear score functions are estimated by applying logistic regression or maximizing the Buffered AUC or AUC with transformed features.

Transforming Features via Cubic Spline Regression

We used a spline transformation to capture the nonlinearity of every feature x_j. We suppose that, for every feature, x_j, an interval $[a, b]$ and a partition

$$a = t_0 < t_1 < \cdots < t_{n-1} < t_n = b,$$

are defined. A spline s on $[a, b]$ is a piecewise nonlinear function

$$s(x) = P_i(x),\ x \in [t_{i-1}, t_i],\ i = 1, ..., n,$$

where P_i is a nonlinear function on $[t_{i-1}, t_i]$, $i = 1, 2, ..., n$. Points $t_i, i = 0, ..., n$ are called knots and are specified by splitting interval $[a, b]$ in subintervals containing (approximately) an equal number of observations (see Ahlberg et al. 2016 for a typical definition of a spline). In this case study, we considered cubic splines, where P_i are cubic polynomials. Cubic splines are commonly used in practice. The fitting procedure can be reduced to convex programming and easily implemented. For every feature, with a PSG package, we maximized the likelihood function for a cubic spline of data. PSG is a general-purpose nonlinear optimization package for solving optimization and statistical problems. Nonlinear transformations, such as $\ln(.)$, $\exp(.)$, or polynomial, are commonly used for transforming features. However, with the splines implemented in PSG, it is possible to perform an optimal transformation rather than using a trial-and-error approach. The mathematical description of one dimensional spline as an argument for the logistic regression likelihood function is described here[7]. The mathematical programming problem for finding optimal parameters of the spline is a convex nonlinear programming problem that can be solved very efficiently with standard nonlinear optimization algorithms. The PSG code for finding splines is in Appendix A. Each feature vector $\boldsymbol{x} \in \mathbb{R}^k$ is transformed via k-separate logistic regression problems, where the j-th logistic regression problem only utilizes the j-th feature x_j to perform a univariate prediction. After estimating splines, say s_j, for every feature x_j, $j = 1, ..., k$, we employ the transformed features $s_j(x_j)$ as the new features of the score function, $S(\boldsymbol{x}) = \sum_{j=1}^{k} w_j s_j(x_j) + b$. This transformation does not affect the complexity of the following optimization problem, such as the maximization of Loglikelihood (1) with logistic regression.

3.3. AUC and Optimization

AUC is a popular criterion in classification. AUC performs quite well for datasets with unbalanced class sizes. It is easy to achieve 99% accuracy on a dataset where 99% of objects are in the same class.

When a score function of a classifier exceeds a threshold, it is considered that the label equals 1; otherwise, the label equals 0. The receiver operating characteristic (ROC) curve is useful for visualizing and evaluating classifiers. The ROC curve is a two-dimensional plot of classifier performance. It is obtained by plotting the true positive rate (TPR) vs. the false positive rate (FPR) for every possible classification threshold.

This section directly maximizes the AUC performance metric. AUC is considered as the objective function for finding a classifier. AUC, by definition, is the area under the ROC curve, see Hanley and McNeil (1982); Bradley (1997); and Fawcett (2006). AUC values range from 0 to 1, since it is a portion of the unit square. A reasonable classifier should have an AUC greater than

[7] Logistics regression likelihood for one dimensional spline is defined in "Example 3. Logarithms Exponents Sum": http://www.aorda.com/html/PSG_Help_HTML/index.html?risk_function_argument.htm.

0.5, because the random guessing generates the diagonal line between $(0,0)$ and $(1,1)$ and gives AUC = 0.5.

AUC has an important statistical property that is equivalent to the Wilcoxon test of ranks, first proposed by Hanley and McNeil (1982). The AUC of a classifier is the probability that the rank for a randomly chosen positive instance is higher than the rank for a randomly chosen negative instance.

Let us denote the score function by $S(x) = \boldsymbol{w}^\top \boldsymbol{s}(x) + b$, where $\boldsymbol{s}(x) := (s_1(x_1), ..., s_k(x_k))^\top$. Note that, when $s_j(x) = x$ for all j, it corresponds to the case where no spline transformation is applied. We suppose that samples with label = 1 have higher scores. By ordering scores $S(\boldsymbol{x}_1), ..., S(\boldsymbol{x}_m)$ and setting a threshold, we get a binary classifier. Further, we provide a probabilistic definition of AUC, see (5). Let L be a loss function vector defined by:

$$L_{ij}(\boldsymbol{w}) = -\boldsymbol{w}^\top (\boldsymbol{s}(\boldsymbol{x}_i) - \boldsymbol{s}(\boldsymbol{x}_j)),\ \ i \in I_1, j\ \in I_0, \tag{2}$$

where $I_0 := \{i \mid y_i = 0\}$ and $I_1 = \{i \mid y_i = 1\}$ denote the index sets with labels -1 and $+1$, respectively. Let us denote by $m_0 := |I_0|$ and $m_1 := |I_1|$ the cardinality of I_0 and I_1. AUC is given by

$$\mathrm{AUC}(\boldsymbol{w}) = \frac{1}{m_1 m_0} \sum_{i \in I_1} \sum_{j \in I_0} \mathbf{1}_{\{L_{ij}(\boldsymbol{w}) \le 0\}} = 1 - \frac{1}{m_1 m_0} \sum_{i \in I_1} \sum_{j \in I_0} \mathbf{1}_{\{L_{ij}(\boldsymbol{w}) > 0\}}, \tag{3}$$

where $\mathbf{1}_C$ is the indicator function of C,

$$\mathbf{1}_C := \begin{cases} 1, & \text{if } C \text{ is true;} \\ 0, & \text{otherwise.} \end{cases}$$

We assume that each $\boldsymbol{x}_i \in I_1$ has probability $\frac{1}{m_1}$ and each $\boldsymbol{x}_j \in I_0$ has probability $\frac{1}{m_0}$. With the cumulative distribution function (CDF) of a real-valued random variable X,

$$F(x) := \Pr\{X \le x\},$$

the probability of exceedance (POE) (see Hoblit 1988) is defined as

$$p_x(X) := \Pr\{X > x\} = 1 - F(x)\,. \tag{4}$$

AUC of the classifier can then be expressed as

$$\mathrm{AUC}(\boldsymbol{w}) = 1 - \Pr\{L(\boldsymbol{w}) > 0\} = 1 - p_0(L(\boldsymbol{w}))\,. \tag{5}$$

We want to maximize the AUC of classifier, i.e., solve the following problem:

$$\max_{\boldsymbol{w}} \quad \mathrm{AUC}(\boldsymbol{w})\,. \tag{6}$$

The PSG package has a precoded probability of exceedance function that can be used for optimization of large datasets (millions of observations) with a standard PC. Since AUC according to Equation (5) can be expressed as the probability of exceedance, we can directly minimize AUC using PSG without writing a special optimization program. The PSG subroutine for minimization of probability is similar to the subroutine for minimizing of quantiles (Value-at-Risk in finance) described in Larsen et al. (2002). Note that the probability of exceedance is an inverse function of the Value-a-Risk.

It can be verified that $\mathrm{AUC}(\lambda \boldsymbol{w}) = \mathrm{AUC}(\boldsymbol{w})$ for any $\lambda > 0$, i.e., the function $\mathrm{AUC}(\boldsymbol{w})$ is positive homogeneous. Therefore, if $\boldsymbol{w}^\star$ is an optimal solution vector of problem (6), then $\lambda \boldsymbol{w}^\star$ is also an optimal solution vector of this problem. We observed that leads to instability of the PSG optimization algorithms, when the optimal point tends to zero.

To make sure that the optimal point is not close to zero, Problem (6) can be equivalently reformulated as follows:

$$\begin{array}{ll} \max\limits_{\boldsymbol{w}} & \mathrm{AUC}(\boldsymbol{w}) \\ \text{s.t.} & \|\boldsymbol{w}\| = 1, \end{array} \tag{7}$$

where $\|\boldsymbol{w}\| := \sqrt{\boldsymbol{w}^\top \boldsymbol{w}}$ denotes the Euclidean norm of a vector $\boldsymbol{w}$. Constraint $\|\boldsymbol{w}\| = 1$ in Problem (7) is nonconvex.

Further, we formulate a problem equivalent to Problem (6), but with a linear constraint (which is a convex constraint). Let us assume that it is known some vector $\boldsymbol{w}^0$ such that $(\boldsymbol{w}^0)^\top \boldsymbol{w}^\star > 0$ for some optimal vector $\boldsymbol{w}^\star$ of Problem (6). Such vector $\boldsymbol{w}^0$ can be a solution of a proxy for Problem (6). For instance, the logistic regression can be considered as a good proxy problem. Let us consider the following AUC maximization problem with a linear constraint:

$$\begin{array}{ll} \max\limits_{\boldsymbol{w}} & \mathrm{AUC}(\boldsymbol{w}) \\ \text{s.t.} & (\boldsymbol{w}^0)^\top (\boldsymbol{w} - \boldsymbol{w}^0) = 0\,. \end{array} \tag{8}$$

The constraint in Problem (8) is imposed to make sure that an optimal solution vector is not equal to 0. Further, we formulate a theorem about the relation of Problems (6) and (8).

Theorem 1. *Let $\boldsymbol{w}^\star$ be an optimal vector of optimization Problem (6), $(\boldsymbol{w}^0)^\top \boldsymbol{w}^\star > 0$ for some vector $\boldsymbol{w}^0$ and $\lambda^\star = \frac{(\boldsymbol{w}^0)^\top \boldsymbol{w}^0}{(\boldsymbol{w}^0)^\top \boldsymbol{w}^\star} > 0$. The following statements are valid:*

(1) $\lambda^\star \boldsymbol{w}^\star$ is an optimal vector of Problems (6) and (8).

(2) In addition, if $\boldsymbol{w}^{\star\star}$ is an optimal solution of Problem (8), then, $\boldsymbol{w}^{\star\star}$ is an optimal solution of Problem (6).

Proof. Let us prove Statement (1) of the theorem. AUC is a positive homogeneous function; therefore, $\mathrm{AUC}(\lambda \boldsymbol{\omega}^\star) = \mathrm{AUC}(\boldsymbol{\omega}^\star)$ for $\lambda > 0$. Consequently, $\mathrm{AUC}(\lambda^\star \boldsymbol{w}^\star) = \mathrm{AUC}(\boldsymbol{w}^\star)$ and $\lambda^\star \boldsymbol{w}^\star$ is an optimal point of Problem (6). Point $\lambda^\star \boldsymbol{w}^\star$ is a feasible point of the constraint in Problem (8). Indeed,

$$(\boldsymbol{w}^0)^\top \left(\lambda^\star \boldsymbol{\omega}^\star - \boldsymbol{w}^0\right) = \lambda^\star (\boldsymbol{w}^0)^\top \boldsymbol{\omega}^\star - (\boldsymbol{w}^0)^\top \boldsymbol{w}^0 = \frac{(\boldsymbol{\omega}^0)^\top \boldsymbol{\omega}^0}{(\boldsymbol{\omega}^0)^\top \boldsymbol{\omega}^\star} (\boldsymbol{w}^0)^\top \boldsymbol{\omega}^\star - (\boldsymbol{w}^0)^\top \boldsymbol{w}^0 = 0\,.$$

Since point $\lambda^\star \boldsymbol{w}^\star$ is feasible for Problem (8) with constraint and optimal for the problem (6) without the constraint, it is also optimal for the problem (8) with constraint. The statement (1) is proved. Let us prove the statement (2) of the theorem. Since $\boldsymbol{w}^{\star\star}$ and $\lambda^\star \boldsymbol{w}^\star$ are optimal solutions of the problem (8), then $\mathrm{AUC}(\boldsymbol{w}^{\star\star}) = \mathrm{AUC}(\lambda^\star \boldsymbol{w}^\star)$. Moreover, $\mathrm{AUC}(\lambda^\star \boldsymbol{w}^\star) = \mathrm{AUC}(\boldsymbol{w}^\star)$, therefore, $\boldsymbol{w}^\star$ is also an optimal solution of problem (6). □

Since AUC equals one minus probability, as shown in Problem (5), AUC maximization Problem (8) can be converted to the following probability minimization problem:

$$\begin{array}{ll} \min\limits_{\boldsymbol{w}} & \Pr\left\{L(\boldsymbol{w}) \geq 0\right\} \\ \text{s.t.} & (\boldsymbol{w}^0)^\top (\boldsymbol{w} - \boldsymbol{w}^0) = 0, \end{array} \tag{9}$$

where L is defined in Problem (2). AUC is discontinuous and nonconvex, which makes it difficult to solve Problem (9) to global optimality. However, PSG has a quite efficient algorithm for minimizing probability with convex constraints. The algorithm used in PSG is similar to the optimization of Value-at-Risk, as described in Larsen et al. (2002). PSG code for solving Problem (9) is included in Appendix A. The main motivation of including Theorem 1 in this section is to explain the reduction of Problem (7) to (9).

3.4. bAUC and Optimization

Several tractable approximation methods have been developed for AUC maximization (see Doucette and Heywood 2008; Miura et al. 2010; and Aiolli 2014). In these methods, AUC was approximated by some surrogate function and this function is maximized.

Norton and Uryasev (2016) suggested a new approach to deal with this complicated problem. They defined Buffered AUC (bAUC) and reduced maximization of bAUC to convex and linear programming problems. The bAUC is the best quasi-concave lower bound of AUC. The bAUC characteristic can be maximized efficiently with convex programming. PSG has a precoded analytical function for bAUC.

The bAUC concept is based on the so-called Buffered Probability of Exceedance (bPOE), defined in Norton and Uryasev (2016), and also in Mafusalov and Uryasev (2018). For references to several papers using bPOE concept in various areas, see Davis and Uryasev (2016); Norton et al. (2017); and Shang et al. (2018). Further, based on one-dimensional minimization representation of the bPOE, Mafusalov et al. (2018) studied statistical properties of empirical estimates of the bPOE. To explain bPOE and bAUC, we give below formal definitions of Value-at-Risk (VaR), Conditional Value-at-Risk (CVaR), and bPOE.

For $\tau \in (0,1)$, the τth quantile $q_\tau(X)$ of X is defined by

$$q_\tau(X) := F^{-1}(\tau) := \min\{x \,|\, F(x) \geq \tau\} = \min\{\, x \,|\, \Pr\{X \leq x\} \geq \tau \,\}\,.$$

In finance, the quantile is known as VaR (see, e.g., Artzner et al. 2002 and Einhorn and Brown 2008).

Rockafellar and Uryasev (2000) considered Conditional ValR (CVaR) as an alternative measure of risk which has better mathematical properties, compared to VaR (see, e.g., Artzner et al. 2002 and Rockafellar and Royset 2018). call CVaR by τ-superquantile for general use outside financial context. CVaR is defined as follows:

$$\bar{q}_\tau(X) := \mathrm{CVaR}_\tau(X) := \min_{\theta \in \mathbb{R}} \left\{ \theta + \frac{1}{1-\tau} \mathrm{E}[X-\theta]^+ \right\},$$

where $[\cdot]^+ = \max\{\cdot, 0\}$. For a continuously distributed random variable X, the CVaR (τ-superquantile) equals a conditional expectation of the tail exceeding the quantile, namely,

$$\bar{q}_\tau(X) = \mathrm{E}[X | X > q_\tau(X)]\,.$$

There are two slightly different variants of bPOE: Upper and Lower bPOEs. In this paper we use Upper bPOE which is defined as

$$\bar{p}_z(X) = \min_{\lambda \geq 0} \mathrm{E}\big[\lambda(X-z)+1\big]^+, \tag{10}$$

for $z \in \mathbb{R}$ such that $\mathrm{E}[X] < z < \sup X$. Formula (10) is considered in Norton and Uryasev (2016) and Mafusalov and Uryasev (2018) as a property of bPOE, but it is convenient to use it as a definition. Further on, Upper bPOE will be called bPOE (without mentioning that it is Upper bPOE). It has been proved in Mafusalov and Uryasev (2018) that bPOE equals $1-\tau$ on the interval $\mathrm{E}[X] < z < \sup X$, where τ is an inverse function of CVaR, i.e., a unique solution of the equation

$$\mathrm{CVaR}_\tau(X) = z,$$

where $\sup X$ is the essential supremum[8] of the random variable X. Therefore, bPOE equals the probability, $1-\tau$, of the tail such that CVaR for this tail is equal to z. The Formula (10) looks quite unusual; the expression (10) does not immediately come across as a probability of some event.

Similar to (5), Norton and Uryasev (2016) defined buffered AUC of a classifier $\boldsymbol{w} \in \mathbb{R}^n$, by

$$\text{bAUC}(\boldsymbol{w}) = 1 - \bar{p}_0(L(\boldsymbol{w}))\,.$$

The maximization of bAUC can be formulated as:

$$\max_{\boldsymbol{w} \in \mathbb{R}^n} \left\{1 - \bar{p}_0(L(\boldsymbol{w}))\right\} = 1 - \min_{\boldsymbol{w} \in \mathbb{R}^n} \bar{p}_0(L(\boldsymbol{w}))\,. \tag{11}$$

Norton and Uryasev (2016) reduced the bAUC maximization (11) to a convex optimization problem:

$$\min_{\boldsymbol{w} \in \mathbb{R}^n} \mathrm{E}\left[L(\boldsymbol{w}) + 1\right]^+, \tag{12}$$

and gave a linear reformulation:

$$\begin{array}{ll} \min\limits_{\boldsymbol{w} \in \mathbb{R}^n, z_{ij} \in \mathbb{R}} & \sum\limits_{i \in I_1} \sum\limits_{j \in I_0} z_{ij} \\ \text{s.t.} & z_{ij} \geq L_{ij}(\boldsymbol{w}) + 1, \quad i \in I_1, j \in I_0, \\ & z_{ij} \geq 0, \quad i \in I_1, j \in I_0. \end{array}$$

It is contrastive that while the AUC maximization (6) through (9) is a nonconvex optimization, the bAUC maximization (12) (with the linear constraint of the form as in Problems (8) or (9)) can be reduced to a linear program. PSG code for minimization of bAUC using bPOE is included in Appendix A. PSG code is based on convex programming and directly utilizes representation (12).

4. Case Study

This section reports a case study, examining the fitting of a simple regression model-based framework to a P2PL application process by applying the methods described above.[9]

4.1. Data Preparation

LendingClub provides open access to sets of loan data since 2007, when the company started operation. The company updates the status of the loans currently listed in downloadable data on a monthly basis and adds new loan data quarterly. The complete loan data are posted for all issued loans. Declined loan data contain the list and details of loan applications that did not meet LendingClub's credit underwriting policy.

The number of reported features changes over time. LendingClub cut open data by 50% in November 2014. At that time, the company removed a half of features of borrowers, as well as all the data of loans with "Policy Code" equal to 2 (new product and not publicly available). That part accounts for 25% of LendingClub's total issuances.

We considered data for three calendar years, 2012, 2013, and 2014. The loan data contain fifty-one features, while the sets of declined loan data have only nine features, as shown below.

- Amount Requested: The total amount requested by the borrower.
- Application Date: The date when the borrower applied for the loan.
- Notes Offered by Prospectus/Loan Title[10]: The loan title or purpose description provided by the borrower.

[8] The essential supremum of the random value X is the smallest number a such that probability of the set $\{X > a\}$ equals zero.

[9] The case study presented in this section (data, codes, and calculation results) is posted at this link http://www.ise.ufl.edu/uryasev/research/testproblems/financial_engineering/%20classification-in-loan-application-process%20/.

- Risk Score: For applications prior to November 5, 2013 the risk score is the borrower's FICO score. For applications after November 5, 2013, the risk score is the borrower's vantage score.
- Debt-To-Income Ratio (DTI): A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
- Zip Code: The first three digits of the zip code provided by the borrower in the loan application.
- State: The state provided by the borrower in the loan application.
- Employment Length: Employment length in years. Values are between 0 and 10 where 0 means "less than one year" and 10 means "ten or more years".
- Policy Code: policy codes 1 and 2 correspond to publicly available and not available new products.

In order to calibrate classification methods, we needed values of features and approved/declined labels for a training set of loan applications. We used only nine features and ignored the descriptive information in "Notes offered by Prospectus" and in "Loan Title". We have not used the feature Policy Code since it equals 0 for every declined loan and 1 for every publicly available loan. Moreover, the information for loans with Policy Code = 2 is not available since November 14, 2014. To make process consistent, we have not considered the Application Date, State, and Zip Code.

Therefore, we considered only four features: Debt-To-Income Ratio, Amount Requested, Risk Score, and Employment Length.

There are a few loan applications with very high values of Debt-To-Income Ratio or Amount Requested, which we regarded as outliers. Following the definition of Tukey (1977), an outlier is a value outside interval

$$[Q_1 - 1.5(Q_3 - Q_1),\ Q_1 + 1.5(Q_3 - Q_1)]\,, \tag{13}$$

where Q_1 and Q_3 are the first and the third quartiles, respectively. We projected outliers to the nearest boundary.

There are many loans with N/A (not available) values for some features. We considered only loans with complete data that do not contain N/A values.

4.2. Numerical Results

We used PSG for conducting numerical experiments. Classification algorithms were evaluated with the AUC criterion.

Spline Transformation of Features

For every feature, we have performed nonlinear spline transformation. Optimal parameters of splines are found by maximizing the logistic regression likelihood of observed labels. Maximization was done with the PSG package (see code in PSG Text format in Appendix A; the code is also available as a MATLAB or R subroutine, see link in Appendix A). We considered five cubic pieces with continuous first, second, and third derivatives at spline nodes. Every spline piece contains about the same number of observations (this is the rule for setting nodes of the spline). Every spline has only eight degrees of freedom ((five pieces) * (four parameters in a cubic polynomial) − (four nodes) * (three constraints) = $5 \times 4 - 4 \times 3 = 20 - 12 = 8$). The dataset contains 380,465 observations in 2012. Therefore, there is no overfitting with the spline optimization procedure, since every spline has only eight degrees of freedom.

Figure 1 shows the spline-transformed DTI feature in 2012. The spline shows the likelihood of approval as function of the DTI feature. We see that loan applications with DTI in the 5–30 range have a relatively high approval likelihood. The graph provides a valuable information for the decision maker about the dependence of approval rate from the feature value.

[10] In some of the sets, this column is Notes Offered by Prospectus, and in others it is Loan Title.

Figure 2 shows the spline transformed Employment Length feature in 2012. The loan applications with Employment Length equal to 0 have a fairly low chance to be approved, as shown by the circled point in Figure 2. The applications with Employment Length greater than 0 have similar chances to be approved. The spline-transformed Risk Score feature is shown in Figure 3. Loan applications with a Risk Score higher around 770 are likely to be approved. It is interesting to observe that applications with the Risk Score higher than 770 have a decreasing chance to be approved. This is a counterintuitive fact. A possible explanation of this fact is that people with very high Risk Score may be overloaded with loans and they may have high DTI (which results in the higher chance of declining the loan).

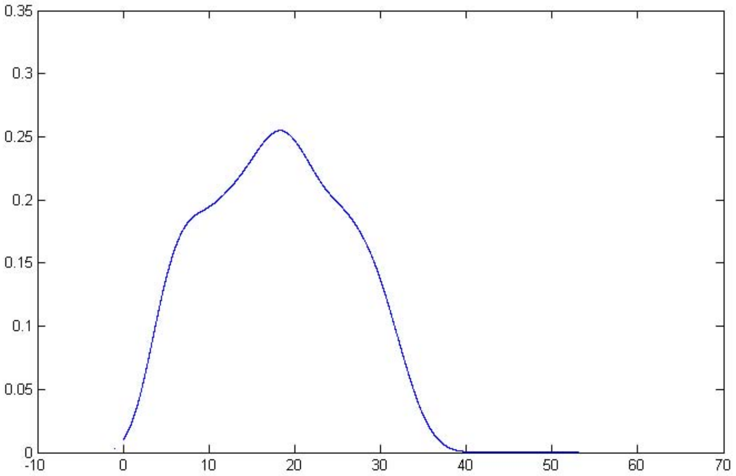

Figure 1. Debt-To-Income (DTI) feature transformed with the cubic spline.

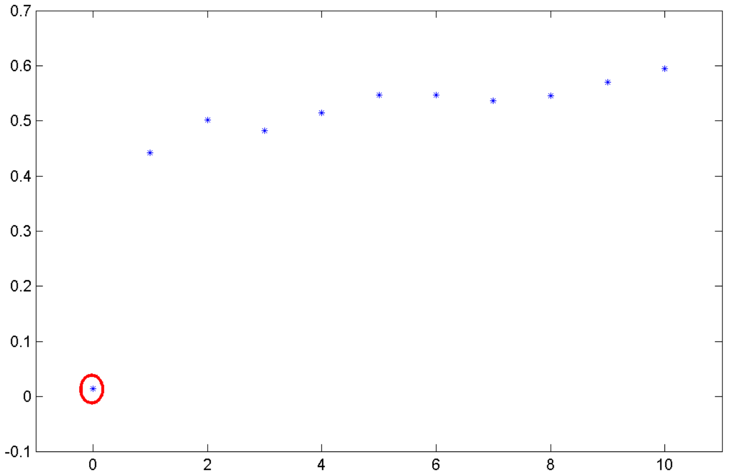

Figure 2. Employment Length feature transformed with the cubic spline. The first observation with distinctively lower value is circled.

We ranked features with the MATLAB 8.1.0.604 (R2013a) subroutine "Rankfeatures" with ROC as the criterion. See MATLAB documentation for the description of the "Rankfeatures" subroutine[11]. For all years (2012, 2013 and 2014), we get the same features ranking, shown in Table 1.

[11] In MATLAB, ROC is the area between the ROC curve and the random classifier slope; link to MATLAB documentation: https://www.mathworks.com/help/bioinfo/ref/rankfeatures.html?searchHighlight=Rankfeatures&s_tid=doc_srchtitle.

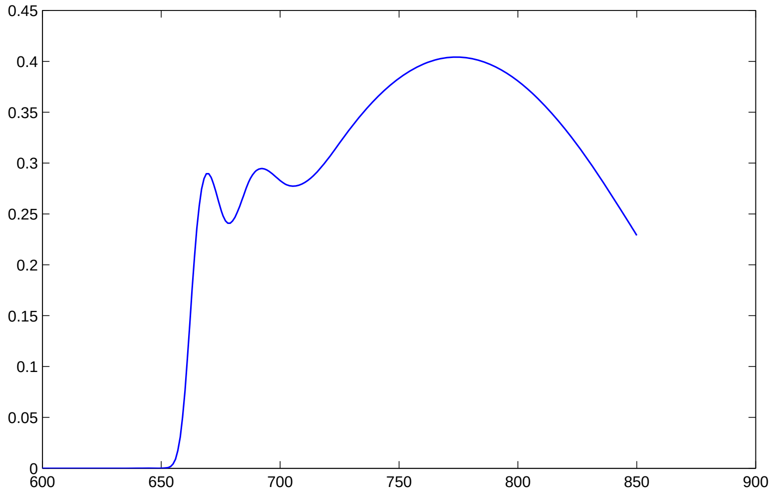

Figure 3. Risk Score feature transformed with the cubic spline.

Table 1. Features Ranking with MATLAB Subroutine "Rankfeatures" (ROC is the criterion).

Ranking	Feature
1	Debt-To-Income Ratio
2	Employment Length
3	Risk Score
4	Amount Requested

4.3. Classification Results

4.3.1. In-Sample Evaluations

This section contains in-sample classification results. First, we conducted the logistic regression with the original and spline transformed features. Data, codes, and calculation results for one instance of logistic regression are in PROBLEM 0 (with original features) and PROBLEM 2 (with spline-transformed features)[12]. Then, we maximized bAUC with the spline-transformed features, as in (11), see PROBLEM 3[11]. Finally, we maximized AUC with the spline-transformed features, where the initial point is set to the solution of the bAUC maximization problem, see PROBLEM 4[11]. PSG codes for Logistic Regression, bAUC maximization, and AUC maximization are in Appendix A.

Model with One Feature: Debt-To-Income Ratio

According to the features ranking, we first used only DTI to perform classification. The resulting AUC values are in Table 2.

The last column shows that with only one feature, we got AUC over 0.72.

Further, we included the Employment Length, which is the second feature in the ranking Table 1.

Model with Two Features: Debt-To-Income Ratio and Employment Length

With two spline transformed features, DTI and Employment Length, we got a fairly high AUC exceeding 0.93 with all considered approaches (see columns 3, 4, and 5 in Table 3).

[12] Data, codes, and calculation results in Text, MATLAB, and R environments. Logistic regression with untransformed features (PROBLEM 0); logistic regression with spline transformed features (PROBLEM 2); maximization of bAUC (PROBLEM 3); maximization of AUC (PROBLEM 4), see link http://www.ise.ufl.edu/uryasev/research/testproblems/financial_engineering/%20classification-in-loan-application-process%20/.

Table 2. Results for models with one feature—DTI.

Year	Logistic Regression without Feature Transformation	Logistic Regression	bAUC Maximization	AUC Maximization
2012	0.535968	0.665422	0.725245	0.725259
2013	0.545397	0.729693	0.729692	0.729707
2014	0.538125	0.724710	0.724709	0.724711

Column 2 shows AUC for logistic regression with the original features without transformation; all other columns show results with spline transformed features. Column 3 contains AUC for logistic regression. Spline transformation significantly improved AUC. Column 4 shows AUC for bAUC maximization (11). Column 5 contains AUC for the AUC maximization Procedure (8).

Table 3. Results for models with two features—DTI and Employment Length.

Year	Logistic Regression without Feature Transformation	Logistic Regression	bAUC Maximization	AUC Maximization
2012	0.905332	0.932214	0.932377	0.932436
2013	0.907454	0.934874	0.935178	0.935191
2014	0.942253	0.973168	0.973479	0.973494

Column 2 shows AUC for logistic regression with the original features without transformation; all other columns show results with spline-transformed features. Column 2 shows that with two features, the logistic regression gives AUC over 0.9. For 2014, AUC maximization with spline transformed features gives AUC over 0.973 (see column 5).

Model with 3 Features: Debt-to-Income Ratio, Employment Length, and Risk Score

We got even higher AUC exceeding 0.95 with three spline transformed features: DTI, Employment Length, and Risk Score (see columns 3, 4, and 5 in Table 4). This table shows that the standard logistic regression (without features transformation) provides quite a high AUC, exceeding 0.93 (see column 2). Further, this result was improved by the spline transformation of features. For 2014, the AUC of logistic regression was improved from 0.937744 to 0.982057.

Table 4. Results for models with three features, DTI, Employment Length, and Risk Score.

Year	Logistic Regression without Feature Transformation	Logistic Regression	bAUC Maximization	AUC Maximization
2012	0.931074	0.954391	0.954673	0.954727
2013	0.934141	0.960136	0.960204	0.960204
2014	0.937744	0.982057	0.982079	0.982097

Column 2 shows AUC for logistic regression with the original features without transformation; all other columns show results with spline transformed features. Logistic regression with spline transformed features gives AUC over 0.95 for all years, see column 3. For 2014, AUC maximization gives AUC over 0.98 (see column 5).

We conducted bAUC maximization and AUC maximization with one, two, and three spline-transformed features (see columns 4 and 5 in Tables 2–4). However, the improvement was insignificant compared to the logistic regression.

We want to mention inconsistency by doing feature transformation with logistic regression and using transformed features for maximizing bAUC and AUC. However, since we got a near-perfect classification (AUC close to 1), transformation of features using bAUC and AUC criteria does not bring additional benefits. Transformation of features using bAUC and AUC criteria are much more complicated problems than the transformation with the logistic regression.

4.3.2. Out-Of-Sample-Evaluations

We also performed out-of-sample validations. We considered the model with three features: DTI, Employment Length, and Risk Score. We have done fourfold cross-validation for 2012. The code for generation of fourfold cross-validation data and solving of four logistic regression problems in PSG text format is in Appendix A (see also "PROBLEM 4"[13]).

The in-sample dataset contains over 380,000 observations. Because of quite the large size of the dataset, there was very little difference between the in-sample and out-of-sample results (in-sample and out-of-sample AUC coincided with four digit precision).

A more meaningful cross validation is to use an optimal solution from previous years for classification in forthcoming years. We took an optimal solution ω obtained by AUC maximization for 2012 and calculated the AUC for 2013. This AUC for 2013 equaled 0.959837, not far from the in-sample AUC obtained by AUC maximization for 2013, which is 0.960204 in column 5, Table 4. We repeated a similar procedure for 2014. We took optimal solution ω obtained by AUC maximization for 2013 and calculated the AUC for 2014. The out-of-sample AUC for 2014 was 0.981841, which is very close to the in-sample AUC obtained by AUC maximization, which is 0.982097 in column 5, Table 4. This out-of-sample verification shows that the same model was used for several years without significant modifications.

5. Conclusions

The objective of this paper was to obtain an understanding of loan classification problems solved by P2PL companies. We considered LendingClub because of the availability of historical data describing the classification decisions. We found that, with three spline-transformed features, Debt-to-Income Ratio, Employment Length, and Risk Score, we obtained an AUC exceeding 0.95, with three considered methods (see columns 3, 4, and 5 in Table 4). In this case, the commonly used logistic regression (without transformation of features) also performed fairly well and delivered AUC exceeding 0.93, see column 2 in Table 4. For some years, e.g., for 2014, the spline transformation of features significantly improved the result (for the logistic regression from 0.937744 to 0.982057). These results indicate that the lending process can be well-approximated with only a few features after adequate nonlinear transformation of those features. We think that the paper provides insight on the loan application process at LendingClub, and suggests an approach that can be used for a similar analysis of other companies.

Regarding methodological improvements, we found that the spline transformation of features may improve the performance of the classification algorithms. We compared three classification optimization procedures: logistic regression, maximization of AUC, and maximization of bAUC. For the considered examples, these methodologies provided quite similar results. Since AUC was used as a criterion, the maximization of AUC outperformed other approaches. However, this outperformance was not significant.

The main conclusion/outcome of this paper is that, although various factors are available for loan selection at LendingClub, these factors are not used. Lending decisions are based only on a few factors that can be processed with standard popular approaches, such as logistic regression. We recommended some improvements, such as the spline transformation of factors, but these improvements do not have significant impact on the overall quality of the classification process. Out-of-sample cross-validation confirmed our conclusions. The case study with data and codes was posted on the website and is available to other researchers.

Author Contributions: All authors contributed to the development and writing of the paper.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

[13] "PROBLEM 4" contains codes for generation of data and solving four-cross validation logistic regression problems in Text, MATLAB, and R environments: http://www.ise.ufl.edu/uryasev/research/testproblems/advanced-statistics/case-sudy-logistic-regression-and-regularized-logistics-regression-applied-to-estimating-the-probability-of-cesarean-section/.

Appendix A. Portfolio Safeguard (PSG) Codes

Appendix A.1. PSG Code for Spline Generation

PSG Text code for obtaining a spline by maximizing the likelihood of logistic regression (see "PROBLEM 1"[14] containing codes in Text, R, and MATLAB formats):

```
maximize
  logexp_sum(spline_sum(matrix_par, matrix_DTI))
```

Matrix `matrix_DTI` is a standard matrix with features and label data, and the matrix `matrix_par` contains spline parameters (we considered five cubic pieces with continuous first, second, and third derivatives at spline nodes; every spline piece contains about the same number of observations).

Appendix A.2. PSG Code for bPOE Minimization (Maximization of bAUC)

PSG Text code for maximization of bAUC (see "PROBLEM 3"[13] containing codes in Text, R, and MATLAB formats):

```
minimize
  bPOE(0,L(matrix_F3_with_label1)-L(matrix_F3_with_label0))
constraint:  = 3.0E+02
  linear(matrix_plane)
```

The matrix `matrix_F3_with_label1` contains rows with features for accepted loans and the `matrix_F3_with_label0` for declined loans. Matrix `matrix_plane` contains coefficients of the linear function in the linear constraint.

Appendix A.3. PSG Code for Probability of Exceedance Minimization (Maximization of AUC)

PSG Text code for maximization of AUC (see "PROBLEM 4"[13] containing codes in Text, R, and MATLAB formats):

```
minimize
  pr_pen(0,L(matrix_F3_with_label1)-L(matrix_F3_with_label0))
constraint:  = 3.0E+02
  linear(matrix_plane)
```

Matrix `matrix_F3_with_label1` contains rows with features for accepted loans and the `matrix_F3_with_label0` for declined loans. Matrix `matrix_plane` contains coefficients of the linear function in the linear constraint.

Appendix A.4. PSG Text Code for Cross-Validation of Logistic Regression

PSG code for generating fourfold cross-validation data and solving four logistic regression problems in PSG Text format (see "PROBLEM 4"[15] containing codes in Text, R, and MATLAB formats).

[14] "PROBLEM 1: problem_Logexp_Sum" contains codes and data for spline transformation in Text, MATLAB and R environments: http://www.ise.ufl.edu/uryasev/research/testproblems/financial_engineering/%20classification-in-loan-application-process%20/.

[15] "PROBLEM 4" contains codes for generation of data and solving four-cross validation logistic regression problems in Text, MATLAB, and R environments: http://www.ise.ufl.edu/uryasev/research/testproblems/advanced-statistics/case-sudy-logistic-regression-and-regularized-logistics-regression-applied-to-estimating-the-probability-of-cesarean-section/.

```
for matrix_fact_in, matrix_fact_out, num = crossvalidation(4, matrix_allscenarios)
  Problem:  problem_num, maximize
  logexp_sum(matrix_fact_in)
end for
```

Matrix `matrix_allscenarios` is a standard matrix with feature and label data.

References

Ahlberg, J. Harold, Edwin Norman Nilson, and Joseph Leonard Walsh. 2016. *The Theory of Splines and Their Applications: Mathematics in Science and Engineering: A Series of Monographs and Textbooks*. Amsterdam: Elsevier, vol. 38.

Aiolli, Fabio. 2014. Convex auc optimization for top-n recommendation with implicit feedback. Paper presented at the 8th ACM Conference on Recommender Systems, Foster City, CA, USA, October 6–10; pp. 293–296.

Artzner, Philippe, Freddy Delbaen, Jean-Marc Eber, David Heath, and Hyejin Ku. 2002. Coherent multiperiod risk measurement. *ETH* Preprint.

Berger, Sven C., and Fabian Gleisner. 2009. Emergence of financial intermediaries in electronic markets: The case of online p2p lending. *BuR Business Research* 2: 39–65. [CrossRef]

Bradley, Andrew P. 1997. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30: 1145–59. [CrossRef]

Chen, Dongyu, Fujun Lai, and Zhangxi Lin. 2014. A trust model for online peer-to-peer lending: A lender's perspective. *Information Technology and Management* 15: 239–54. [CrossRef]

Collier, Benjamin C., and Robert Hampshire. 2010. Sending mixed signals: Multilevel reputation effects in peer-to-peer lending markets. Paper presented at the 2010 ACM Conference on Computer Supported Cooperative Work, Hangzhou, China, March 19–23. pp. 197–206.

Davis, Justin R., and Stan Uryasev. 2016. Analysis of tropical storm damage using buffered probability of exceedance. *Natural Hazards* 83: 465–83. [CrossRef]

Ding, Jie, Jinbo Huang, Yong Li, and Meichen Meng. 2018. Is there an effective reputation mechanism in peer-to-peer lending? Evidence from China. *Finance Research Letters*. [CrossRef]

Doucette, John, and Malcolm I. Heywood. 2008. Gp classification under imbalanced data sets: Active sub-sampling and auc approximation. In *European Conference on Genetic Programming*. Berlin/Heidelberg: Springer, pp. 266–77.

Einhorn, David, and Aaron Brown. 2008. Private profits and socialized risk. *Global Association of Risk Professionals* 42: 10–26.

Emekter, Riza, Yanbin Tu, Benjamas Jirasakuldech, and Min Lu. 2015. Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending. *Applied Economics* 47: 54–70. [CrossRef]

Fawcett, Tom. 2006. An introduction to roc analysis. *Pattern Recognition Letters* 27: 861–74. [CrossRef]

Freedman, David A. 2009. *Statistical Models: Theory and Practice*. Cambridge: Cambridge University Press.

Habermann, Shelby J. 1979. *Analysis of Qualitative Data: Introductory Topics*. Cambridge: Academic Press.

Hanley, James A., and Barbara J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 143: 29–36. [CrossRef] [PubMed]

Hoblit, Frederic M. 1988. *Gust Loads on Aircraft: Concepts and Applications*. Reston: American Institute of Aeronautics and Astronautics.

Hosmer, David W., Jr., Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*. Hoboken: John Wiley & Sons, vol. 398.

Hulme, Michael K., and Collette Wright. 2006. Internet based social lending: Past, present and future. *Social Futures Observatory* 11: 1–115.

Iyer, Rajkamal, Asim Ijaz Khwaja, Erzo FP Luttmer, and Kelly Shue. 2009. *Screening in New Credit Markets: Can Individual Lenders Infer Borrower Creditworthiness in Peer-To-Peer Lending?* Rochester: SSRN.

Jiang, Cuiqing, Zhao Wang, Ruiya Wang, and Yong Ding. 2018. Loan default prediction by combining soft information extracted from descriptive text in online peer-to-peer lending. *Annals of Operations Research* 266: 511–29. [CrossRef]

Lai, Linda S. L., and Efraim Turban. 2008. Groups formation and operations in the web 2.0 environment and social networks. *Group Decision and Negotiation* 17: 387–402. [CrossRef]

Larsen, Nicklas, Helmut Mausser, and Stanislav Uryasev. 2002. Algorithms for optimization of value-at-risk. In *Financial Engineering, E-Commerce and Supply Chain*. Berlin/Heidelberg: Springer, pp. 19–46.

Lending Academy. 2010. Available online: http://www.lendacademy.com/ (accessed on 1 November 2018).

LendingClub. 2006. Available online: https://www.lendingclub.com/ (accessed on 1 November 2018).

Lin, Mingfeng. 2009. Peer-to-peer lending: An empirical study. *AMCIS 2009 Doctoral Consortium* 17: 1–7.

Lin, Mingfeng, Nagpurnanand R. Prabhala, and Siva Viswanathan. 2013. Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science* 59: 17–35. [CrossRef]

Ma, Ben-jiang, Zheng-long Zhou, and Feng-ying Hu. 2017. Pricing mechanisms in the online peer-to-peer lending market. *Electronic Commerce Research and Applications* 26: 119–30. [CrossRef]

Mafusalov, Alexander, Alexander Shapiro, and Stan Uryasev. 2018. Estimation and asymptotics for buffered probability of exceedance. *European Journal of Operational Research* 270: 826–36. [CrossRef]

Mafusalov, Alexander, and Stan Uryasev. 2018. Buffered probability of exceedance: Mathematical properties and optimization. *SIAM Journal on Optimization* 28: 1077–103. [CrossRef]

Mi, Jackson J., Tianxiao Hu, and Luke Deer. 2018. User data can tell defaulters in p2p lending. *Annals of Data Science* 5: 59–67. [CrossRef]

Miura, Kakeru, Satoshi Yamashita, and Shinto Eguchi. 2010. Area under the curve maximization method in credit scoring. *The Journal of Risk Model Validation* 4: 3. [CrossRef]

Norton, Matthew, Alexander Mafusalov, and Stan Uryasev. 2017. Soft margin support vector classification as buffered probability minimization. *The Journal of Machine Learning Research* 18: 2285–327.

Norton, Matthew, and Stan Uryasev. 2016. Maximization of auc and buffered auc in binary classification. *Mathematical Programming* 1–38. [CrossRef]

Puro, Lauri, Jeffrey E. Teich, Hannele Wallenius, and Jyrki Wallenius. 2010. Borrower decision aid for people-to-people lending. *Decision Support Systems* 49: 52–60. [CrossRef]

Rockafellar, Ralph Tyrrell, and Johannes O. Royset. 2018. Superquantile/cvar risk measures: Second-order theory. *Annals of Operations Research* 262: 3–28. [CrossRef]

Rockafellar, Ralph Tyrrell, and Stanislav Uryasev. 2000.Optimization of conditional value-at-risk. *Journal of Risk* 2: 21–42. [CrossRef]

Shang, Danjue, Victor Kuzmenko, and Stan Uryasev. 2018. Cash flow matching with risks controlled by buffered probability of exceedance and conditional value-at-risk. *Annals of Operations Research* 260: 501–14. [CrossRef]

Smith, Andrew M. 1999. Sec cease-and-desist orders. *Administrative Law Review* 51: 1197.

Tsai, Kevin, Sivagami Ramiah, and Sudhanshu Singh. 2014. *Peer Lending Risk Predictor.* Stanford University CS229. Stanford: Stanford University.

Tukey, John W. 1977. *Exploratory Data Analysis*. Reading: Sage, vol. 2.

Wang, Hui, Martina Greiner, and Jay E Aronson. 2009. People-to-people lending: The emerging e-commerce transformation of a financial market. In *Value Creation in E-Business Management*. Berlin/Heidelberg: Springer, pp. 182–95.

Wu, Jinghua, and Yun Xu. 2011. A decision support system for borrower's loan in p2p lending. *JCP* 6: 1183–90. [CrossRef]

Yu, Haihong, MengHan Dan, Qingguo Ma, and Jia Jin. 2018. They all do it, will you? Event-related potential evidence of herding behavior in online peer-to-peer lending. *Neuroscience Letters* 681: 1–5. [CrossRef] [PubMed]