Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Predicting Peer-to-Peer Loan Rates Using Bayesian Non-Linear Regression

Zsolt Bitvai
University of Sheffield
Sheffield, United Kingdom
z.bitvai@shef.ac.uk

Trevor Cohn
University of Melbourne
Melbourne, Australia
t.cohn@unimelb.edu.au

Abstract

Peer-to-peer lending is a new highly liquid market for debt, which is rapidly growing in popularity. Here we consider modelling market rates, developing a non-linear Gaussian Process regression method which incorporates both structured data and unstructured text from the loan application. We show that the peer-to-peer market is predictable, and identify a small set of key factors with high predictive power. Our approach outperforms baseline methods for predicting market rates, and generates substantial profit in a trading simulation.

Introduction

Peer-to-peer (P2P) lending is an emerging market that facilitates lending between individuals and small-medium sized companies. One characteristic of these markets is that they exhibit highly non-linear behavior, with substantial noise. Temporal dynamics can be observed as well due to varying loan terms and changing market conditions. Last, social lending involves significant quantities of user generated text which is an integral part of the loan application process. All this key information can be incorporated in the Gaussian Process framework which can handle non-linear mapping functions. These latent functions are well suited to handle noise, and the posterior of the model supports exploitation of market inefficiencies via automatic trading.

In this paper we present a Gaussian Process regression model for predicting market loan rates, which infers an underlying latent function linking the attributes of each loan and its rate, while accounting for the considerable noise in the data. This models not just the rates for unseen loans, but also the uncertainty of the predictive posterior, which allows for full Bayesian inference, namely, to detect mispriced loans on the market, and quantify expected profits that could be realized by buying and reselling loan parts at correct market value.

The Gaussian Process is a state of the art machine learning framework which allows for rich modelling of several aspects of the data. We incorporate temporal variation, structured data, and unstructured text using a kernel combination technique that modulates between addition and multiplication. We find that this modulation beats a model with purely additive or multiplicative kernel combination.

Each loan application contains rich information about the borrower, third party credit scoring, textual information, such as a questions and answers section as well as a cover letter from the borrower explaining their background and the purpose of the loan. There are around 1,000 – 1,500 words of English text per document. In order to harness the semantic content from this text, we use LDA topic modelling and TF/IDF scoring of terms. We find that our model outperforms baselines and other well known linear and non-linear learning algorithms such as support vector regression.

Related Work

Text regression is the problem of predicting a quantifiable real world phenomenon from unstructured text inputs. This has seen diverse applications including the prediction of stock volatility from company reviews (Kogan et al. 2009), market movements from news (Lerman et al. 2008), movie revenues from critic reviews (Joshi et al. 2010) and political opinion polls from social media posts (Lampos, Preotiuc-Pietro, and Cohn 2013). Almost exclusively, these approaches have used linear predictive models, which aids model interpretability, nonetheless these approaches are inherently unable to handle non-linearities (one notable exception is Wang and Hua (2014), who modelled stock volatity using a non-linear regression model). In particular, this is likely to be problematic for financial market data, which is known to be noisy and highly non-linear (Atiya 2001). In this paper we propose a Gaussian Process model with non-linear kernels for modelling P2P lending rates, which we show improves significantly over linear modelling. We retain model interpretability through automatic relevance determination (ARD) and learned kernel importance weightings, which identify the relative importance of features and feature groups.

In peer to peer lending, rich data is presented to investors in order to allow them to make informed decisions. Having a way to combine several disparate features is important because in our modelling problems, in common with many others, we have rich structured data in addition to unstructured text. For example, Joshi et al. (2010) predicted movie revenues not just from the text of critic reviews, but also used

Copyright © 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

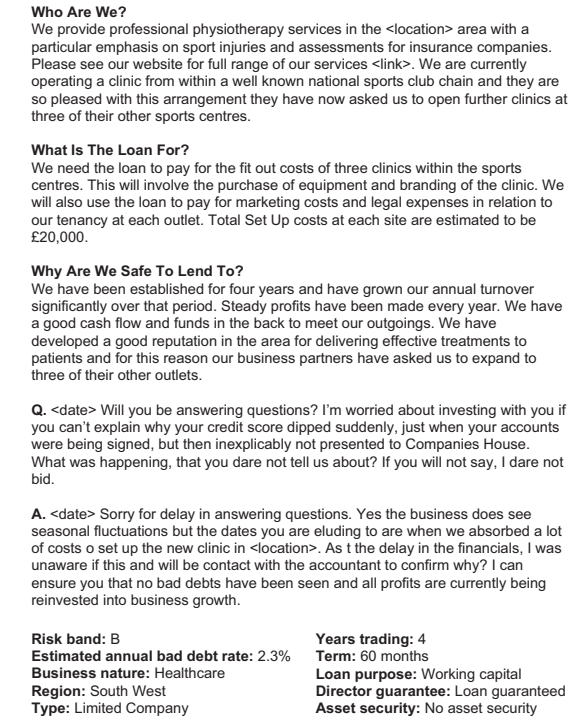
data such as genre and rating of the movie, showing that the combination of both inputs provides the best performance. The complementarity of structured and text data was also observed in financial market modelling (Lerman et al. 2008) and predicting Ebay auction outcomes (Resnick and Zeckhauser 2002). For this reason, we seek to combine text with structured data in modelling P2P loan rates, proposing several kernel combination methods combined with Bayesian model selection.

Besides unstructured text and structured numerical data, additional information can be included by examining the time dependency of data. Financial market dynamics are known to change considerably with time, such as for Ebay auctions in Ghani and Simmons (2004). Previous work has adopted sliding window or time bucketing for training in order to prevent past data from unduly influencing future predictions (Kogan et al. 2009; Lampos, Preotiuc-Pietro, and Cohn 2013; Wang and Hua 2014). In contrast we take a more principled approach by explicitly modelling time using several kernels over temporal features, such as the date and the time left on a loan. These allow modelling of smoothly varying changes with time, while learning the importance and rate of change of each temporal factor.

A remaining question is how to represent text. Many models make a "bag-of-words" assumption, a simplification which facilitates fast training and scaling to large datasets. However, text has much deeper semantic meaning which incorporates its context in the document and the corpus. There have been mixed results in prior work on text regression when attempting to use deeper representations for text. Kogan et al. (2009) found that TF/IDF alone does fairly well even without additional text processing. Similarly, Joshi et al. (2010) found that part of speech tagging and dependency relations did not contribute to performance improvement. In this work, we use TF/IDF scores and extract semantic meaning with LDA topic modelling (Blei, Ng, and Jordan 2003), which can automatically capture latent semantic information in the document that would not be accessible directly from the bag-of-words. Although TF/IDF scores have been used for text regression in the past, in our model we also add the LDA topic distribution to the feature set.

P2P Loan Markets

Peer to peer lending is a new industry that enables people to lend and borrow money from each other directly, akin to crowdfunding. When applying for new loans on Funding Circle[1], users have to fill out an application form and they are subject to credit scoring. Then there is a period of conversation between potential lenders and the borrower in the form of Questions and Answers. Lenders can competitively offer money to the borrower for the period of two weeks in the form of an auction. Once the auction is closed, the borrower can choose to accept the offers made with the lowest interest rate or reject the auction. If accepted, the loan becomes live and investors can trade their loan parts with one another in a secondary market, where sales often represent a premium or discount relative to the original investment. The secondary market rate is the quantity that we seek to model, based on the rich details of the loan along with temporal factors. In the end, a loan either defaults, where investors lose their money, is repaid early if the borrower no longer needs the funds, or matures until its term length and expires naturally. Investors are repaid their capital plus interest on a monthly basis.



Who Are We?
We provide professional physiotherapy services in the <location> area with a particular emphasis on sport injuries and assessments for insurance companies. Please see our website for full range of our services <link>. We are currently operating a clinic from within a well known national sports club chain and they are so pleased with this arrangement they have now asked us to open further clinics at three of their other sports centres.

What Is The Loan For?
We need the loan to pay for the fit out costs of three clinics within the sports centres. This will involve the purchase of equipment and branding of the clinic. We will also use the loan to pay for marketing costs and legal expenses in relation to our tenancy at each outlet. Total Set Up costs at each site are estimated to be £20,000.

Why Are We Safe To Lend To?
We have been established for four years and have grown our annual turnover significantly over that period. Steady profits have been made every year. We have a good cash flow and funds in the back to meet our outgoings. We have developed a good reputation in the area for delivering effective treatments to patients and for this reason our business partners have asked us to expand to three of their other outlets.

Q. <date> Will you be answering questions? I'm worried about investing with you if you can't explain why your credit score dipped suddenly, just when your accounts were being signed, but then inexplicably not presented to Companies House. What was happening, that you dare not tell us about? If you will not say, I dare not bid.

A. <date> Sorry for delay in answering questions. Yes the business does see seasonal fluctuations but the dates you are eluding to are when we absorbed a lot of costs o set up the new clinic in <location>. As t the delay in the financials, I was unaware if this and will be contact with the accountant to confirm why? I can ensure you that no bad debts have been seen and all profits are currently being reinvested into business growth.

Risk band: B
Estimated annual bad debt rate: 2.3%
Business nature: Healthcare
Region: South West
Type: Limited Company
Years trading: 4
Term: 60 months
Loan purpose: Working capital
Director guarantee: Loan guaranteed
Asset security: No asset security

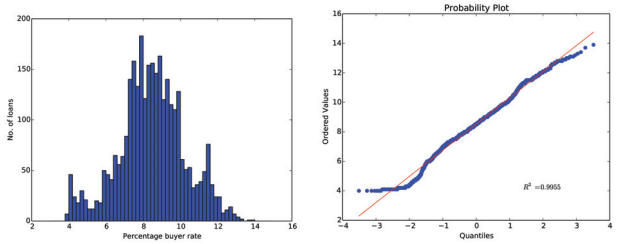
Figure 1: Excerpt from a loan auction (sic, redacted).

An excerpt from the loan auction page can be seen in Figure 1. The main auction page contains information about the requested amount by the borrower, the term length of the loan, 60 months in the excerpt, how long the borrower has been trading in years, i.e. 4 years, and the target rate at which they would like to borrow, which is tied to the risk band, "B" in the example. Additional data visible in the excerpt includes the nature of the business (Healthcare), which roughly corresponds to industry, the geographical region the business operates in (South West of the United Kingdom), the type of business (limited company), the risk band which includes the estimated annual rate of bad debt (2.3%), the purpose of the loan (Working capital), and whether there is asset security (no) or a director's guarantee (yes).

The borrower is subject to credit scoring, which takes the form of a year of monthly measurements by external agencies. The current credit score is also calculated relative to all businesses, relative to the industry the business is in, relative to businesses with similar size and with similar age as the borrower's business.

Apart from the credit history, a detailed profit and loss statement of the applicant is provided. This is the annual filed management accounts of the company for the last few

[1] http://fundingcircle.com



(a) (b)

Figure 2: Distribution of buyer rates, showing a histogram over the rates (a), and a normal probability plot (b). The red line in (b) illustrates perfect normality, and the data points are plotted in blue.

years. Additional information includes the amount of outstanding loans, the amount of overdraft, and the payment performance for the company and the industry.

Each loan application contains textual information on the purpose of the loan and the borrower's business. This is covered by the "Who Are We," the "What Is The Loan For," and the "Why Are We Safe To Lend To" sections, as shown in Figure 1. This information is filled out by the borrower when they make a loan application. Lenders and borrowers can also communicate with each other through a Questions and Answers section where lenders can ask questions, which the borrower can answer.

After bidding finishes, the closing rate and the accepted bid spread of the loan is available. The closing rate is the weighted average rate of all the accepted bids. Finally, we record time dependent data, such as when the auction closed at, as well as the date we scraped the secondary market prices, and the time difference between these two. We have gathered around 3,500 loan applications between 2013 and 2014. There are around 1,000 - 1,500 words in 50-70 sentences with an average of around 6 question and answer pairs and 3 additional text fields per loan request.

The target of the model is to predict the secondary market ongoing buyer rate of loans, which is defined as the top of the ask order book, i.e. the highest rate currently on offer for the loan among all the offers. As Funding Circle does not provide a bid side of the order book, or historical trade data, this is the closest indicator of market judgment, where a lower rate means less risky. This can deviate significantly from the rate the loan closed at, as many market participants only buy exclusively from the secondary market and do not participate in the primary market. Another target can be the markup applied to loans.

Examining the distribution of secondary market buyer rates in Figure 2 we see that it passes the nearly normal condition. The dips in the -2 and +3 quantiles give evidence of risk aversion and risk seeking behavioral biases (Kahneman and Tversky 1979), where loans with the lowest and highest rates are unusually in high demand. Below quantile -2, we see the effect of a hard minimum limit imposed on the market place at 4% with some values below that due to rounding error on the market server. Additional spikes can be observed at the minimum bid rates, i.e. quantile 1.5 - 11.5%, where a market service called "autobid" automatically buys loan parts with a 0% markup. The target ranges between 4-15%.

Model

Here we use a Gaussian Process (GP) regression model, a Bayesian kernelized extension of linear regression which models Gaussian covariance between pairs of datapoints (Williams and Rasmussen 2006). A GP prior defines a latent set of random values over an input space (here feature vectors), where any finite values of variables follow a multi-variate Gaussian distribution (here noise-corrected loan rates). GP regression is formulated as

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$
$$y \sim \mathcal{N}(f(\mathbf{x}), \sigma_n^2)$$

where $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$ is the mean function and $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$ is the covariance function, which are both user defined functions. The observations y are assumed to be noise-corrupted versions of the latent process, where σ_n^2 is the variance of the Gaussian noise. We follow standard practice by letting $m = 0$, while we consider a suite of covariance functions designed to work with the various aspects of our input (structured, temporal, text). Given a labelled training sample, $(\mathbf{X}, \mathbf{y})$, consisting of $(\mathbf{x}, y)$ pairs, we can express the Bayesian posterior, and integrate out the latent function values, $\mathbf{f} = f(\mathbf{X})$, resulting in an analytic formulation for the marginal likelihood, $p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})d\mathbf{f}$. Although this involves the inversion of the covariance matrix with complexity $O(n^3)$ where n is the number of training instances, scaling to larger datasets can be accommodated using approximation methods with linear or sub-linear time and space complexity in n (Williams and Rasmussen 2006; Hensman, Fusi, and Lawrence 2013). Maximizing the log marginal likelihood using gradient ascent allows for the noise parameter σ_n^2, as well as other model hyperparameters (parameterizing k, described below) to be optimized.

The choice of the covariance function, k, varies depending on the application and the nature of the features. Here we use the rational quadratic covariance function (Williams and Rasmussen 2006) defined as

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma_d^2 \left\{ 1 + \sum_{d=1}^{D} \frac{(x_{id} - x_{jd})^2}{2\alpha l^2} \right\}^{-\alpha} \quad (1)$$

where D is the number of features, and σ_d^2, l and α are the kernel hyper-parameters, named amplitude, lengthscale and power respectively. This covariance function is equivalent to an infinite sum over radial basis function (RBF) kernels of different lengthscales; setting $\alpha = \infty$ recovers the RBF. Accordingly, this kernel is appropriate if patterns are present in the data occurring over different lengthscales, e.g., global correlations combined with local smoothness, such as long term trend in loan rates and short term variations in trading patterns. For some features we use Automatic Relevance Determination (Neal 1995) with the above covariance function, which uses a different scale parameter for each input

dimension, i.e., the l^2 term in (1) is replaced with l_d^2. Accordingly, the ARD identifies the relevance of features based on their relative lengthscales, with shorter scales indicating more rapid fluctuations in the output with the feature, and accordingly higher importance.

Since we have different types of features extracted from the data, the question remains as to how best to combine these. Different types of features can be integrated by adding, multiplying or by convolving other valid covariance functions over the data (Williams and Rasmussen 2006). For this reason, we define the addmul kernel operator that modulates between kernel addition and multiplication,

$$\mathrm{addmul}(k_1, ..., k_n) = \prod_{i=1}^{n}(k_i + b_i) \tag{2}$$

where k is an input kernel (defined over input pairs, omitted for clarity) and b is a bias constant. Expanding this formulation reveals a weighted sum of multiplicative combinations of the input kernels. Collectively the bias, $\mathbf{b}$, and scaling hyperparameters (σ_d^2 for each kernel, see Eq. 1) determine the weighting of each kernel combination term. The intuition behind the modulated kernel is that is we do not want to hardcode either multiplication or addition over the kernels but rather naturally find a balance between these operators, fit by optimizing the Bayesian marginal likelihood. This comes down to the difference between arithmetic and geometric averaging, namely that while the arithmetic (addition) average smooths out component variation, the geometric (multiplication) allows each component to significantly affect the overall result, e.g., a component kernel close to zero will lead to overall ∼zero covariance. Having all views agreeing is desirable, however this is unlikely to happen often in practice, hence modulating the geometric mean to consider combinations of fewer components is a more beneficial strategy. This means the modulated kernel can express standard kernel addition and multiplication, as well as products over subsets.

Text is often high dimensional, owing to a large vocabulary size, which complicates the use of ARD with per-word features due to twin issues of optimization convergence and overfitting. Therefore, we apply the standard rational quadratic kernel over the textual feature groups (LDA and TF/IDF) and the ARD rational quadratic over the time and numeric features groups. These kernels are combined with the addmul operator,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{addmul}(k_t, k_n, k_l, k_f) \tag{3}$$

where K is the covariance between two $\mathbf{x}$ loan instances, k_t is the kernel over the time features, k_n is the kernel over the numeric feature group, k_l is the kernel over the LDA topic distributions and k_f is the kernel over the TF/IDF scores. Together the bias terms, each kernel's σ_d, l and α values, and the Gaussian noise variance σ_n^2 form the model hyperparameters, which are fit by optimizing the marginal likelihood.

After training, which involves fitting the model hyperparameters to maximize the marginal likelihood of a supervised training sample, we seek to make predictions on unseen data. Under GP regression, the predictive posterior over the loan rate, y_*, for test point $\mathbf{x}_*$, is another Gaussian, $p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$ which can be analytically computed (see Williams and Rasmussen (2006) for details). We use the predictive mean for evaluating predictive error rates, but as a secondary evaluation consider its use in a trading scenario. In a trading context, assume we buy a loan part with observed rate $\hat{y}$ and attempt to resell it later at rate y. For maximum profit, we want to maximize the difference between these two rates, $U(y|\hat{y}) = \hat{y} - y$, which quantifies the extent to which the loan is underpriced. Assuming that the market rate is distributed according to the GP posterior, we can quantify the expected profit as

$$\begin{aligned}\mathbb{E}[U(y|\hat{y})] &= \Pr(y_* \leq y)U(y|\hat{y}) \\ &= \Phi(y|\hat{\mu}, \hat{\sigma}^2)U(y|\hat{y})\end{aligned} \tag{4}$$

where $\Phi(y|\hat{\mu}, \hat{\sigma}^2)$ is the Gaussian cumulative density function of the posterior. The optimizing y is then found using gradient ascent of (4), which we use to price the loan part for sale. Alternatively, the expected profit can be employed to identify the best trading opportunities across several different candidate loans. The optimization can be performed with respect to markup as well, which can be derived from the rate and other loan properties.

Features

The feature groups used in this experiment are shown in Table 1, we now elaborate on each of these features.

Structured Features

It is hypothesized that key loan information such the requested amount, the term length of the loan and the closing interest rate affect the buyer rates of the loans. Historical credit scores are likewise considered an important factor since they influence the probability of default. A number of crucial financial indicators can be extracted from the annual filed management accounts of the companies for the last few years, and based on this information, a trader can infer the health of the business. We extract the Altman financial ratios (Atiya 2001) as an indicator of bankruptcy probability for small and medium sized companies, similar to those listed on Funding Circle. These are are defined as:

1. Working capital / Total assets
2. Retained earnings / Total assets
3. Earnings before interest and taxes / Total assets
4. Market capitalization / Total debt
5. Sales / Total assets

The more money a borrower requests, the less time they have been trading, and the longer they need the money, the more riskier the application is perceived. The estimated annual bad debt is Funding Circle's assessment of the business's health, and therefore included.

The closing rate and the bid spread gives indication as to how other investors have judged the loan application, and will likely influence the buyer rate as it is these investors that will sell the loan parts initially. It is likely that further

Table 1: Input feature groups

Numeric
closing rate
min accepted bid
max accepted bid
Altman ratio 1-5
term length
trading years
amount requested
bad debt rate
credit history
relative score

Time
auction date
scrape date
elapsed time
LDA
100 topic freqs
TF-IDF
$\sim$ 3000 word scores

information exists in the depths of bids and the investor identities, and thus we perform text analysis both with and without these identities.

Other potential modelling improvements could come from using the categorical data to support multi task learning, e.g., using coregionalization or hierarchical kernels, to uncover richer relationships in the data, such as explicit modelling of different sectors or individual investors.

The auction closing date and scrape date give information about changing market conditions at various points in time. The elapsed time since the closing date indicates how long the borrower has been repaying the loan, which could influence the probability of future default. This is a basic form of survival analysis, where loans that survive for a longer time are considered less risky and thus have a lower buyer rate.

Unstructured Features

The Questions and Answers section is a form of crowd due diligence, the purpose of which is to provide additional assessment of the loan application by the wisdom of the crowds. An example of this can be seen in Figure 1. In addition, many of the people asking questions also trade on the secondary market, giving insight into their valuation of the loan. The description section gives information about the motives of the borrower and their background, which affects the probability that they will not be able to repay the money.

TF/IDF, short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus (Rajaraman and Ullman 2011). It is often used as a weighting factor in information retrieval and text mining. The TF/IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others. For example, in Table 2 we can see that "aircraft," "roof" and "nightclub" are highly specific and important terms to the documents, but "explain" and "small" are much more general terms and accordingly we expect them to be of much less use for our application of loan rate modelling. The presence of specific lenders in an auction is likely to correlate with the secondary market loan rates, which raises the possibility of modelling individual investors

Table 2: Top/bottom 5 TF/IDF words from selected documents (left) and top 10 words from LDA topics (right).

aircraft	roof	nightclub
machin	scaffold	bar
aerospac	fibr	floor
program	min	tabl
agreement	tile	music
explain	explain	run
gener	larg	larg
small	go	much
level	small	circl
next	get	number

construct	hous	independ
sector	beer	currenc
score	brew	scottish
tax	pub	affect
deliv	northern	repair
debt	per	paid
contractor	week	hous
tender	breweri	run
recent	weekli	vend
may	craft	car

as part of our analysis. Therefore, we include the investor usernames who bid on the loan or pose questions in the Q&A section as words, which are used in computing the TF/IDF scores and LDA topics, described below.

Before computing word scores, the text is pre-processed by filtering out punctuation, and replacing numerical values with a placeholder tag. Then, word tokens are stemmed using Porter stemmer (Porter 1980) and stop words are removed. Next, words with a count of less than 20 and the top 100 words are removed from the dataset. Finally, TF/IDF scores are calculated for each remaining word type, resulting in an approx. 3,000 dimensional vector space representation of each loan document.

Latent Dirichlet Allocation (LDA; Blei, Ng, and Jordan 2003) is a soft document and word clustering method based on a latent document representation. It posits that each document is a mixture of a small number of topics where each word's creation is attributable to one of the document's topics.

We use LDA to infer for each loan document a distribution over 100 latent topics, where each topic is a scalar value between 0 and 1, summing to 1 across all topics. The vector of topic assignments are then used as features of the loan, which is incorporated into the model using a rational quadratic kernel, as described above. Since questions and answers are fairly sparse, we pool these fields together with the borrower description. However, it may be useful to separate these or align terms from the questions to the answers. LDA captures the semantic content of loan requests which reflects the themes lenders and borrowers are concerned about when contemplating investment decisions. In Table 2 we can see that Scottish independence is one such issue. The presence of certain investors in the topics may be indicative of latent investment preferences, for example, for certain industries or risk bands.

Results

We fit a Gaussian Process over the normalized data with 4-fold crossvalidation and measure the root mean square error and marginal likelihood. The marginal likelihood can be employed for Bayesian model selection to choose between competing hypotheses (Williams and Rasmussen 2006). It includes both a data fit term and a model complexity penalty, thereby penalizing overly permissive models. Moreover, this is evaluated solely on the training data, and thereby avoids

Table 3: Main results, showing baseline (top) versus modelling approaches (bottom). The columns report cross-validation RMSE and training marginal likelihood.

Description	RMSE	log $\mathcal{L}$
Baseline Average Buyer Rate	1.780	-
Baseline Closing Rate	1.352	-
Ridge Regression	1.004	-
SVR Linear	0.970	-
SVR RBF	0.994	-
GP Linear	0.738	-1116
GP RBF	0.607	-847
GP RatQuad	**0.596**	-811

Table 4: Comparative results of the GP RatQuad model with different kernel configurations.

Description	RMSE	log $\mathcal{L}$
Operator		
Add	0.689	-1127
Prod	0.699	-1382
Addmul	**0.596**	-811
Features		
Time	1.599	-3069
Numeric	0.771	-1379
Numeric Top 6	0.764	-1373
Time+Numeric Top 6	0.602	-842
LDA	1.225	-2516
TFIDF	0.972	-2328
TFIDF+LDA	0.936	-2146
Numeric+LDA+TFIDF	0.680	-1084
Time+Numeric+LDA+TFIDF	**0.596**	-811

the need for heldout cross-validation. As baselines we include predicting the closing rate of the auction as the secondary market buyer rate, and predicting the average buyer rate in the training set. In addition to GP regression, support vector regression (SVR) with linear and RBF kernels, and ridge regression are evaluated for comparison. The hyperparameters are tuned using grid search over a logarithmic scale.

The results in Table 3 show that the model confidently beats the baselines (1.352), linear regression (1.004), as well as SVR (0.970), with a root mean square error of **0.596**. Note that the scale of the output variable ranges between 4-15 as can be seen in Figure 2. The best GP kernel is the rational quadratic, which beats the linear and RBF kernels. This suggests that there are multiple overlapping patterns in the dataset that the RatQuad kernel can identify, but RBF cannot due to lack of flexibility. Table 4 shows comparative results of the GP RatQuad model. Note the close correspondence between maximizing marginal likelihood and minimizing test RMSE and that model selection will select the model with the lowest error (0.596). The addmul operator outperforms both addition and multiplication as a way to combine kernels, suggesting that modulation between addition and multiplication is effective. Textual features are shown

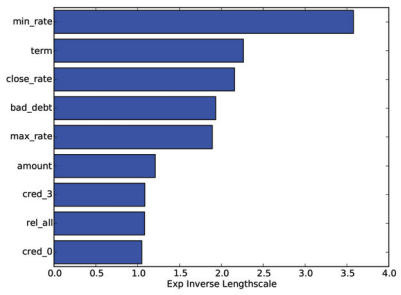


Figure 3: Most important numerical features, showing the log inverse length scale hyperparameters. Larger values mean greater contribution to the covariance.

to contain important information, with models using only text features (e.g., TFIDF+LDA) outperforming the baselines and competing SVR approaches. When combined with structured numeric features, they result in further improvements. The temporal kernel on its own is not very effective, but combining the time kernel with the other kernels resulted in significant improvements. Last, it can be seen that LDA and TF/IDF features have identified complementary signals in the data to the numerical and temporal components. When combining all features, the best model accuracy is achieved.

With respect to the importance of numerical features in predicting buyer rate, as can be seen in Figure 3, we find that the most important features are the closing rate (`close_rate`) and the spread of the accepted bids, where the minimum accepted bid (`min_rate`) is weighted more than the maximum (`max_rate`). Other important features are the term length of the loan (`term`) and the bad debt ratio of the risk band for the loan (`bad_debt`). The amount requested is slightly less important (`amount`). In the credit score features, we see the average company score across all companies dominates (`rel_all`), which is an indicator of overall economic environment. With regard to monthly credit history, we find that the most important ones are the current score (`cred_0`) and the one three months ago (`cred_3`). In fact, we find that ARD is effective at feature selection, and rerunning the numeric model with only the top 6 features selected results in slightly improved predictive accuracy (see `Numeric Top 6` in Table 4). In addition, it appears that there is some amount of information content overlap between the various feature groups, as the key numeric features combined with time do fairly well. Traditional important financial metrics such as credit scores, profit and loss accounts and borrower personal statements however only bear marginal influence on market rates, which questions the usefulness of fundamental analysis in this regard. In the text based models, we find that the key indicators are the identities of the individual investors committed to certain loan auctions, which suggests the model is able to uncover latent correlations of investment patterns and preferences of the users.

Next, we simulate trading on the secondary market over a unit of time. This acts as a proof of concept to demonstrate

how the model predictions could potentially be exploited. Profitability of arbitrage opportunities can be crudely calculated by assuming lack of short selling, no transaction costs and that market rates are distributed according to our model posterior. Note, if we had a time series of market depth for each loan, that would allow for more sound empirical evaluation here. The simulation starts by buying each loan part at the listed rate, and attempting to sell it above a new market rate using Eq. 4. If a sale occurs, which we can estimate using the model posterior, we record the profit as the difference between the buy and sell rate. Otherwise, we hold on to the loan part and record zero profit. Using this procedure, we measure the compound return over the approx. 3,500 test loan parts starting with a unit of capital, resulting in an end capital of R = 98.35. This compares to a result of 0.005 for the baseline method of randomly pricing the loan parts under the posterior.

Conclusions

We have observed that P2P markets are predictable and we have shown a simple way to model them with a GP and maximize expected profits in a trading scenario. Via model and feature selection we have identified the variables that explain moving secondary market rates on the Funding Circle market place. Text based features have successfully uncovered preferences of individual investors, while traditionally important metrics such as profit and loss accounts, credit history and borrower personal statements have shown to be less informative. In the future, we plan to model other targets such as loan markup and the likelihood of default, deeper use of the categorical data using e.g., multi task learning (Cohn and Specia 2013) and hierarchical kernels (Hensman, Lawrence, and Rattray 2013) and experiment with the Student's-T likelihood which is more resilient to outliers.

Acknowledgements

Dr Cohn is the recipient of an Australian Research Council Future Fellowship (project number FT130101105).

References

Atiya, A. F. 2001. Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks* 12(4):929–935.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Cohn, T., and Specia, L. 2013. Modelling annotator bias with multi-task gaussian processes: An application to machine translation quality estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

Ghani, R., and Simmons, H. 2004. Predicting the end-price of online auctions. In *Proceedings of the International Workshop on Data Mining and Adaptive Modelling Methods for Economics and Management*.

Hensman, J.; Fusi, N.; and Lawrence, N. D. 2013. Gaussian processes for big data. *Proceddings of the Conference on Uncertainty in Artificial Intelligence*.

Hensman, J.; Lawrence, N. D.; and Rattray, M. 2013. Hierarchical bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BioMed Central Bioinformatics* 14(1):252.

Joshi, M.; Das, D.; Gimpel, K.; and Smith, N. A. 2010. Movie reviews and revenues: An experiment in text regression. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 293–296.

Kahneman, D., and Tversky, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society* 263–291.

Kogan, S.; Levin, D.; Routledge, B. R.; Sagi, J. S.; and Smith, N. A. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 272–280.

Lampos, V.; Preotiuc-Pietro, D.; and Cohn, T. 2013. A user-centric model of voting intention from social media. In *Proc 51st Annual Meeting of the Association for Computational Linguistics*, 993–1003.

Lerman, K.; Gilder, A.; Dredze, M.; and Pereira, F. 2008. Reading the markets: Forecasting public opinion of political candidates by news analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, 473–480.

Neal, R. M. 1995. *Bayesian learning for neural networks*. Ph.D. Dissertation, University of Toronto.

Porter, M. F. 1980. An algorithm for suffix stripping. *Program: electronic library and information systems* 14(3):130–137.

Rajaraman, A., and Ullman, J. D. 2011. *Mining of massive datasets*. Cambridge University Press.

Resnick, P., and Zeckhauser, R. 2002. Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. *Advances in Applied Microeconomics* 11:127–157.

Wang, W. Y., and Hua, Z. 2014. A semiparametric gaussian copula regression model for predicting financial risks from earnings calls. In *Proceedings of the 52nd Annual Meeting on Association for Computational Linguistics*.

Williams, C. K., and Rasmussen, C. E. 2006. Gaussian processes for machine learning. *MIT Press* 2(3):4.