Topological Analysis of Credit Data: Preliminary Findings

James Cooper[1][0000−0002−2962−2829], Peter Mitic[2,3][0000−0002−9845−4435], Gesine Reinert[4][0000−0002−0363−9470], and Tadas Temčinas[4][0000−0001−8618−0812]

[1] Santander US, Boston MA 02109, USA
[2] Department of Computer Science, University College London, UK
[3] Santander UK, London, UK
[4] Department of Statistics, University of Oxford, UK

Abstract. There is plenty of room for improvement in credit risk prediction. Intuitively, similar customers should have similar credit risk. Capturing this similarity is often carried out using Euclidean distances between customer features and predicting credit default via logistic regression. Here we explore the use of topological data analysis for describing this similarity. In particular, persistent homology algorithms provide summaries of point clouds which relate to their topology. This approach has been shown to be useful in many applications but to the best of our knowledge, applying topological data analysis to prediction of credit risk is novel. We develop a pipeline which is based on the topological analysis of neighbourhoods of customers, with the neighbourhoods given through a geometric network construction. Using two data sets from the *Lending Club* we find a modest signal; the results have high variance, but they could be seen as indication that including such topological features could improve credit risk prediction when used as additional explanatory variable in a logistic regression.

Keywords: Credit Risk · Topological Data Analysis · Barcode · Landscape · Logistic Regression

1 Introduction

The context for this paper is bank unsecured lending. The bank makes its lending decisions using customer details (or *features*) such as the customer's employment status, income, expenditure, current total unsecured debt, previous credit record, and other third party and derived features (such as debt-to-income ratio). No single factor, nor combination of factors, is a sure-fire predictor of 'success' in repaying the loan. Bank predictions mostly succeed, but there is room for improvement. A better understanding of the probability of default can make even risky customers profitable. The heuristic behind credit risk prediction is that customers with "similar" features should be associated with similar risk of default. Finding useful measures of similarity is an ongoing issue in credit risk forecasting.

Here we investigate the use of Topological Data Analysis (*TDA*), which embraces the topological relationship between a single datum and other data. As such, there is a direct expression in topological terms of the phrase "If customer A paid back a loan, and customer B looks like customer A, then customer B should also pay back a loan". There are many applications of *TDA* in a financial context (e.g. [10]) but to the best of our knowledge, it has not been applied to credit risk modelling. Outside finance, there have been successes of classifiers based on *TDA* in many areas such as neuron spike data [17], cancer prognosis [5], and image classification [2]. This work follows previous discussions on financial credit-worthiness using data from the Lending Club (*LC - https://www.lendingclub.com/*), a US-based personal loan organisation.

To validate the pipeline, in the absence of related work for credit scoring, we apply and compare the pipeline to oncology data [8] [5]. Using that data set, Wu and Hargreaves [19] carried out a related study, using logistic regression enhanced with TDA features. We set a high bar. For a training set, we use a set which achieves very high accuracy based on logistic regression alone, rather than a random sample. Hence, achieving similar results when including topological information, which is what we find, is promising.

The main contributions of this paper are, first, the finding that there can be a topological signal in credit risk data, and second, a pipeline for including topological summaries in credit scoring.

Nomenclature: In this paper the acronyms *LR* and *TDA* refer, respectively, to the *logistic regression* curve-fitting method, and the application of topological concepts to data analysis. We use the acronym *LR+TDA* to refer to a logistic regression calculation that incorporates topological components.

2 Literature Review: Credit Scoring

We concentrate on a summary of previous work on credit analysis. Work on *TDA* will be discussed in Section 3. Credit scoring was developed in the early 1940s in an attempt to control credit risk. In 1941 Durand [9] developed a 'scorecard' formula based on factors such as a customer's salary, age, sex, credit history, and occupation. Variants on that system are still used today. More factors have been introduced, and the weights attached to each factor have matured. A notable example is the U.S. *FICO* credit score [6], originally formulated in 1956 by the *Fair Isaac Corporation.* The idea of *nearest neighbours*, a key *TDA* component, was first applied by Chatterjee in 1970 [7]. In 1980, Ohlson [14] applied multivariate discriminator analysis (MDA) in an early probabilistic model of corporate bankruptcy prediction; similar probabilistic models have since been applied to retail contexts. An early application of a Logistic Regression (*logit*) model was formulated by Wigginton in 1980 [18]. *Logit* models were an advance on *MDA* models by removing the *MDA* requirement for characteristics to be

[5] https://archive.ics.uci.edu/ml/datasets/heart+disease

[6] https://www.fico.com/

drawn from a multivariate normal distribution. Further techniques, including genetic algorithms, neural networks and decision trees are discussed in [11].

3 Concepts in Topological Data Analysis

Here we briefly introduce some concepts from topological data analysis which are useful for this paper. More detailed introductions can be found, for example, in [15] and [6]. Following [6], we start with a set of points P in Euclidean space $\mathbb{R}^N$ with Euclidean distance, the norm $\|\cdot\|_2$. For each point $p \in P$ we define its $\epsilon-$neighbourhood $N_\epsilon(p)$ as the ball $B_\epsilon(p)$ of radius ϵ around p.

3.1 The Vietoris-Rips Filtration

For increasing ϵ, the balls around points will increasingly overlap, and eventually there will be no holes left between the balls. The process of increasing ϵ is used to construct characteristics of the point cloud P. First we create a sequence of so-called *Vietoris-Rips simplicial complexes* $VR_\epsilon(P)$. We start with a graph having vertices P and edges (p_0, p_1) for all pairs of points p_0, p_1 in P with distance $\|p_0 - p_1\|_2 \leq \epsilon$. We then set

$$VR_\epsilon(P) = \bigcup_{l\geq 0} VR_\epsilon(P)_l, \quad VR_\epsilon(P)_l = \{(p_0, \ldots, p_l) \mid \ \|p_i - p_j\|_2 \leq \epsilon \text{ for all } i, j\}.$$

Here $VR_\epsilon(P)_l$ can be viewed as a list of all l-simplices of the complex $VR_\epsilon(P)$. This construction is called *Vietoris-Rips filtration.* For this object we can calculate so-called *homology groups.* We shall only use the first two homology groups, which are easy to describe intuitively: The dimension-0 homology counts the number of connected components, and the dimension-1 homology counts the number of holes. The process as ϵ increases is illustrated in Figure 1, reproduced from [6].

3.2 Barcode

A *barcode* is a visual representation of point cloud connectivity as ϵ increases. A non-zero element (connected component for dim 0, and hole for dim 1) that first appears at $\epsilon = \epsilon_b$ and vanishes at $\epsilon = \epsilon_d$ is represented by the interval (or *bar*) $(\epsilon_b, \epsilon_d]$.

The process described in Subsection 3.1 takes a finite set of points $P \subseteq \mathbb{R}^N$ as an input and for every homological dimension $k \leq N$ outputs a barcode $\mathcal{B}(P, k)$, which can be represented as a multiset of intervals of the form $\left\{(\epsilon_b(i), \epsilon_d(i)]\right\}_{i=1}^{m_k}$. An important feature of barcodes is that they are directly comparable between each other by many different metrics that are stable, in the sense that a small perturbation in the input point-cloud leads to only a small perturbation of the barcode, as measured by the metric. For an overview of stability results, see [16, Ch. 3].

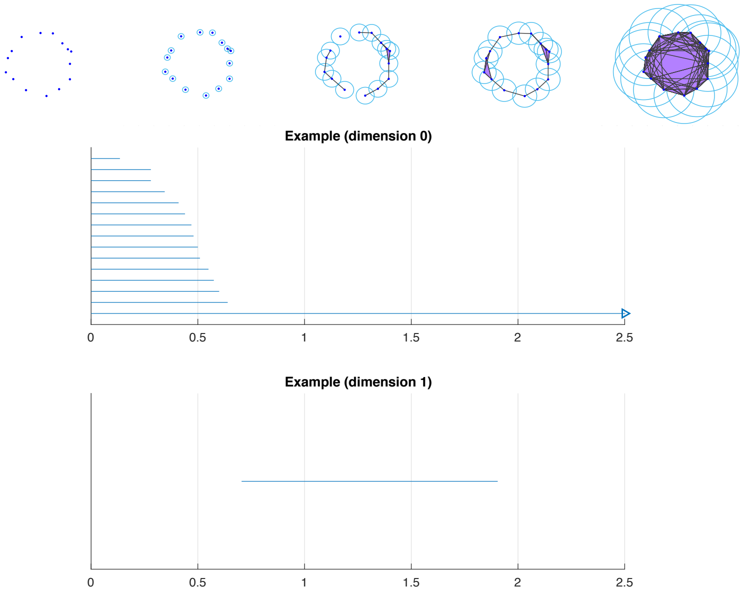

Fig. 1. An example of a Vietoris-Rips filtration and the corresponding barcodes in dimension 0 and dimension 1. Figure 1 from [6].

3.3 The Wasserstein Metric

Here we describe the 1-Wasserstein metric, used here to measure distance between barcodes. Let $\mathcal{B}_1, \mathcal{B}_2$ be two barcodes such that every homological feature, like a connected component or a hole, has to persist for a finite time. Then the 1-Wasserstein metric is defined as:

$$W_1(\mathcal{B}_1, \mathcal{B}_2) = \inf_{\gamma \in \mathit{\Gamma}} \sum_{(\epsilon_b, \epsilon_d] \in \mathcal{B}_1 \cup \mathit{\Delta}} \left\| (\epsilon_b, \epsilon_d] - \gamma((\epsilon_b, \epsilon_d]) \right\|_\infty , \tag{1}$$

where $\mathit{\Gamma}$ is the set of bijections from $\mathcal{B}_1 \cup \mathit{\Delta}$ to $\mathcal{B}_2 \cup \mathit{\Delta}$. Here $\mathit{\Delta} = \{(x, \infty) : x \in \mathbb{R}\}$, which we use for technical reasons to take care of the case when the barcodes do not have the same number of bars. The distance between two intervals $(a, b]$ and $(a', b']$ is defined as $\|(a, b] - (a', b']\|_\infty := \max(|a - a'|, |b - b'|)$. Note that in theory the 1-Wasserstein distance can be infinite.

3.4 Landscape

An alternative to the barcode encoding of topological information is a *persistence landscape* [4]. A *landscape* is derived from the set of m_k birth-death points at dimension k: $\left\{(\epsilon_b(i), \epsilon_d(i)]\right\}_{i=1}^{m_k}$ of a *barcode*. With $\epsilon_b(i)$ plotted against $\epsilon_d(i)$, a piecewise continuous function, termed a *landscape*, can be defined to model the extremities of the plot. In practice, we use a discretised form of this function, at some predefined resolution. Since we use dimension 0 and 1 homologies, we get two *landscapes* per point cloud. The precise mathematical definition of *persistence landscapes* is beyond the scope of this paper, and we refer the interested reader to [4].

We would like to highlight a few important properties of landscapes. They are effectively vectorised expressions of barcodes at the predefined resolution (we used 500 reference points). All landscapes for a point cloud can be expressed in terms of the same resolution. That makes them easy to use in existing statistical and machine learning techniques, especially as they and their corresponding metrics are fast to calculate compared to Wasserstein distances (given in Equation (1)). Second, they are also stable: a small perturbation of the point cloud will only result in a small perturbation of the landscape [4, Section 5]. Third, as real-valued functions, we can leverage different statistical techniques (such as adding, averaging) when analysing them.

Figure 2 shows an example of a typical *barcode* and the corresponding *landscapes* used in our analysis.

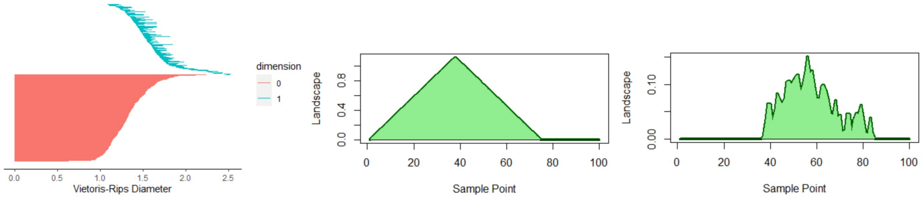

Fig. 2. Left to right: typical *LC* Barcode, with Landscapes at dimensions 0 and 1.

4 Application Pipeline to Credit Data

Topology is essentially a tool to analyse a structure, not an individual node. Therefore, a preliminary step is to impose a structure on each node by defining a list of neighbouring nodes. Here we represent credit data as a network between customers, with two customers connected by an edge if their feature vectors are close in Euclidean space. With such a network representation we then calculate the barcodes and landscapes for each individual neighbourhood. We use these topological summaries as additional measures of similarity in a logistic regression.

In more detail, we assume that a credit data set is available with N customers, and each customer has n features associated with it. The features are normalised to $[0, 1]$. A *node* with index i, n_i, comprises a set of n normalised features x_{ij}, and an binary outcome Y_i representing successful repayment of a loan or not (1 or 0 respectively); $n_i = \Big\{\{x_{i,1}, x_{i,2}, \ldots, x_{i,n}\}, Y_i\Big\}$ with $x_i \in [0, 1], Y_i \in \{0, 1\}$. A collection of N nodes, C_N, constitutes the *point cloud* which is the basis of the credit risk *TDA* analysis; $C_N = \left\{n_i\right\}_{i=1}^{N}$. The neighbourhood of a target node n_i is a set of other nodes $\{n_j, j \neq i\}$ that are 'close to' n_i, according to some 'closeness' criterion. To construct edges in this network, one could use a *k-nearest-neighbours* (*kNN*) algorithm with a Euclidean distance metric. However, we have found that a *k-nearest-neighbours* approach is very slow to calculate. Instead, we have used the *kd-Tree* approximation [1]. In the *kd-Tree* algorithm, the sample space is bisected, and the Euclidean nearest neighbours algorithm is applied only within the subspace containing n_i. The containing subspace is repeatedly bisected in the same way until a required neighbourhood size is reached. This method is considerably faster than *kNN*, but can omit 'near' nodes that are in the 'wrong' subspace. In *R*, the *kd-Tree* algorithm is implemented in the *RANN* package, which is a wrapper for the *ANN C++* library.

Once the network is constructed, each node n_i (customer) is assigned its neighbourhood as the subgraph induced by the nodes $n_j, j \neq i$ such that there is an edge between n_i and n_j in the network.

4.1 Homologies for neighbourhoods

The central step when applying *TDA* in any context is the calculation of persistent homologies for each node. Persistent homologies are calculated using the neighbourhoods at dimensions 0 and 1. The dimension 2 components of the barcodes for the data used are minor, and we assume that they have a negligible effect on the results. The homologies are then used to calculate outcome predictors. The *Wasserstein* predictor is based on the 1-Wasserstein distance metric, Equation (1) (see e.g. [12]), which accounts for geometry of the *barcodes* being compared. The Wasserstein predictor for a Test node reports the proportion of Training nodes in its Wasserstein neighbourhood who repay the loan, separately for dimension 0 and 1. The *Landscape* predictor is derived from a persistence diagram *Landscape* construct [3]. It reports the proportion of Training nodes in the Landscape neighbourhoods who repay the loan, again separately for dimensions 0 and 1. With a *Wasserstein* and a Landscape predictor for each dimension, there is a total of four predictors. Finally, the predictors are used as a classifier for a set of Test nodes. The predicted outcomes are compared to the actual test outcomes.

4.2 Detailed LR+TDA Credit Risk algorithm

The *LR+TDA* pipeline details are described in the algorithm in this section. The first step, generation of a single Training set plus multiple Test sets, was designed to identify an "optimal" training set using the original data only. Specifically, this training set yields maximum *Accuracy* in multiple trials. The *high accuracy* training set presents a severe test for the case where the original data are augmented by *TDA* predictors. Consequently, results for *TDA*-augmented data that are close to those of the original data are sought.

1. Sample data
   (a) Draw 100 random samples (balanced 50:50 between defaulted and not), partition each into Training/Test sets in a ratio 2/3 : 1/3
   (b) Run LR calculations for each Training/Test pair, then choose the training set that yields the highest LR Accuracy.
   (c) Generate further similar Test sets
2. Calculate neighbourhoods using the $kd - Tree$ algorithm, such that the resulting number of neighbours is at least the square root of the number of Training nodes, subject to the two conditions below.
   (a) The neighbourhood for each Training node is derived from all other Training nodes
   (b) The neighbourhood for each Test node is derived from all Training nodes
3. Calculate Training homologies, relative to Training nodes, separately for outcomes $Y = 0$ and $Y = 1$
4. Calculate Test homologies, also relative to Training nodes, separately for outcomes $Y = 0$ and $Y = 1$
5. Find 1-Wasserstein distances between neighbourhoods:
   (a) For all Training nodes, calculate the 1-Wasserstein distances relative to other Training nodes, at dimensions 0 and 1. Select the M (we used $M = 25$) nearest nodes
   (b) For all Test nodes, calculate the 1-Wasserstein distances relative to all Training nodes, at dimensions 0 and 1. Select the M nearest nodes
6. Calculate Wasserstein predictors
   (a) For all Training Wasserstein neighbourhoods, calculate the proportion of those corresponding successful predictions at dimensions 0 and 1
   (b) For each Test node, calculate the proportion of Training nodes in its Wasserstein neighbourhood with Outcome = 1
7. Find least squares distances between Landscapes of neighbourhoods
   (a) Calculate landscapes for all nodes
   (b) Calculate least squares distance metrics for all Training landscapes, relative to other Training landscapes, at dimensions 0 and 1. Select the M least (we used $M = 25$)
   (c) Calculate least squares distance metrics for all Test landscapes, relative to all Training landscapes, at dimensions 0 and 1. Select the M least
8. Calculate Landscape predictors
   (a) For all Training Landscape neighbourhoods, calculate the proportion of the corresponding successful predictions at dimensions 0 and 1
   (b) For each Test node, calculate the proportion of Training nodes in its Landscape neighbourhood with Outcome $Y = 1$
9. Logistic regressions
   (a) Augment the original data with additional features in separate calculations: the *Wasserstein* predictors alone, the *Landscape* predictors alone, then both together
   (b) Do the LR calculations for all combinations.
10. Assess signal detection on Test data using a *Voting* algorithm: If predictions for *Wasserstein* and *Landscape* agree at dimension 0, accept the common dimension 0 result. Otherwise, accept the mean prediction at dimension 1.

The Wasserstein calculations impose a severe restriction on the use of *TDA* with large data sets. The time taken to do them makes it infeasible to process more than 1000 nodes. Therefore we draw random samples from the data, and repeat each complete *LR+TDA* calculation 10 times. We have found that using more than 10 repetitions reduces standard deviations of the results only minimally.

5 Results

5.1 Data

Two data sets, sourced from the *Lending Club*, give details of unsecured loan applications, either defaulted or not, for the period 2007-14. The first, *LC-B* has 188000 nodes and 56 features. Its default rate is 15.7%. Predictions for this data set have proved to be particularly difficult [13]. The second, *LC-A* has 42500 nodes and 25 features, with a default rate 15.1%. [7] Both default rates are high for a European bank. The *LC* results are compared with the results of two alternative data sets. The first, *Japan Credit Screening Data* is a small credit data set (690 nodes with 13 features), known to yield a high *LR* accuracy. Its source is the *UCI* repository. [8] The second is the *Cleveland oncology* data [8], also sourced from the *UCI* repository, included to compare our results with those of an independent *TDA* study on the same data [19] in order to validate our procedure as there is no comparable study on credit risk data available. All implementations were programmed in *R* on an Intel i7 quad core processor with 64GB RAM, using in particular, packages *TDA* and *TDAStats*.

5.2 Exploratory Classification Analysis

First we assess whether there is a topological signal in the data when not including logistic regression (step 10 of Algorithm 4.2). The results are collated in Table 1. We detect a signal in sensitivity and precision. For both *LC* data sets, the signal is fairly weak and the overall success rate is not significantly different from a random guess. However, even a small increase in precision may create an improvement in assessing credit risk. For the *Japan* data, no sampling was required and hence no standard deviation is reported; the signal is strong.

5.3 Sampling Results

Step 1 of Algorithm 4.2 gave details of how the Training set and Test sets were generated. A total of ten Test sets were generated, and *LR+TDA* calculations were carried out on all Training-Test combinations. With samples of size 500 (333 Training, 167 Test), each complete *LR+TDA* calculation took about 50 minutes to complete. The *Japan* and *Oncology* Training data sets were selected in the

[7] Neither is now available from the *Lending Club* website.

[8] https://archive.ics.uci.edu/ml/datasets/Japanese+Credit+Screening

Table 1. Success Indicators, *LC* and *Japan* data, exploratory analysis

Data		% Accuracy	Sensitivity	Specificity	Precision
LC-B	Mean	51.74	0.529	0.506	0.516
LC-B	SD	3.40	0.03	0.06	0.03
LC-A	Mean	54.19	0.55	0.54	0.54
LC-A	SD	3.86	0.06	0.04	0.04
Japan	Mean	74.35	0.827	0.641	0.739
Japan	SD	n/a	n/a	n/a	n/a

same way, except that all data were used, so that only one test set is necessary. The results are shown in Table 2. The column headers are: *None* = Original data only, *Wass* = Original data augmented by Wasserstein dimension 0 and 1 predictors, *Land* = Original data augmented by Landscape dimension 0 and 1 predictors, and *Both* = Original data augmented by both types of predictor.

Table 2. Success Indicators, *LC-B* and *LC-A*, sample size 500, 10 runs each: *Japan*, all data 1 run

		Mean				**SD**			
Data	**Metric**	**None**	**Wass**	**Land**	**Both**	**None**	**Wass**	**Land**	**Both**
LC-B	% Accuracy	57.01	56.29	57.07	54.85	4.72	5.00	4.17	3.97
LC-B	Sensitivity	0.77	0.78	0.78	0.79	0.16	0.15	0.16	0.16
LC-B	Specificity	0.37	0.35	0.36	0.31	0.2	0.19	0.18	0.18
LC-B	Precision	0.55	0.55	0.55	0.53	0.04	0.04	0.03	0.03
LC-B	AUC	0.63	0.63	0.63	0.62	0.03	0.03	0.03	0.03
LC-A	% Accuracy	60.66	59.82	59.58	59.88	4	3.28	4.13	4.73
LC-A	Sensitivity	0.52	0.49	0.51	0.52	0.11	0.11	0.13	0.11
LC-A	Specificity	0.69	0.7	0.68	0.67	0.1	0.1	0.11	0.11
LC-A	Precision	0.63	0.63	0.62	0.62	0.05	0.05	0.05	0.06
LC-A	AUC	0.63	0.63	0.63	0.63	0.05	0.05	0.05	0.05
Japan	% Accuracy	90.0	83.48	90.0	83.48	n/a	n/a	n/a	n/a
Japan	Sensitivity	0.87	0.76	0.87	0.93	n/a	n/a	n/a	n/a
Japan	Specificity	0.93	0.93	0.93	0.72	n/a	n/a	n/a	n/a
Japan	Precision	0.94	0.93	0.94	0.8	n/a	n/a	n/a	n/a
Japan	AUC	0.96	0.9	0.96	0.96	n/a	n/a	n/a	n/a

The results with *TDA* indicate parity with those without, with some modest improvements in the means of some metrics. The *Both* cases show that Wasserstein and Landscape should not, in general, be used together. We assume that they provide contradictory indications of success. It is encouraging that improvements (i.e. increases) in success metric means are usually matched by improvements (i.e decreases) in success metric standard deviations. Comparison of results with those from the *Oncology* and *Japan Credit* data provide further evidence that results including *TDA* features are comparable to those without. In particular, the independent study [19] using the same oncology data con-

Table 3. Comparison of *Cleveland* oncology analyses [8], with and without *TDA*, showing figures for best configurations found.

Publication	**Accuracy without TDA**	**Accuracy with TDA**	**Sensitivity**	**Specificity**	**Precision**
Wu [19]	88.14	89.83	0.89	0.91	0.90
This paper	92.08	92.08	0.93	0.90	0.93

firms our 90% accuracy figure, and also our figures for sensitivity, specificity and precision. Our result comparisons with and without *TDA*-augmented data are stringent because the training sets are optimised for maximum *LR Accuracy*. *LR+TDA* can compete, just!

6 Discussion

Three key considerations when implementing a credit decisioning model are:

1. Can customer creditworthiness be scored in (near) real time?
2. Is the scoring "fair", can it proved to be so, and is it explainable?
3. How can a model be expressed in terms of a scorecard?

Each of these factors make linear modelling (after sigmoid transformation, logistic regression) appealing and, indeed, logistic regression has become the *de facto* method of credit analysis. Methods have been optimized by individual banks over the past 30 years. Much of this optimization arises from data collected for credit scoring. With recent progress in explainability of models (see the DALEX and LIME packages in R) there is a path to work with Regulators to help them become comfortable with the features being engineered.

Here we have explored features which arise from *TDA*. A *TDA* aspiration is to capture features that a linear model cannot detect. We find indications for a subtle signal in the *TDA* summaries which we chose, but the case is perhaps not compelling. We have tried minor modifications (such as weighting importance of neighbour scores by proximity; combining dimensions of *TDA* results in different ways) but without significant change of predictive power.

Other improvements may be possible, such as extending the voting procedure to include nodes outside the neighbourhood. However, in our view, a more fundamental change is needed. We propose a focus on how to expand the number of records we may use to train models, as well as introducing novel features whose structure is not revealed by linear methods. The next steps are for us are to see if combining *TDA* predictions based on different training samples can provide more insight, and to find if Filtration-derived neighbourhoods provide a viable way forward.

Moreover, the high standard deviations in the *LR+TDA* predictions may be an indication of substantial heterogeneity in the data which is not modelled. Similar to personalised medicine, we shall try to identify subgroups of customers for which predictions can be reached with substantially smaller variance than for the entire population.

Acknowledgements GR was supported in part by EPSRC grants EP/T018445/1 and EP/R018472/1. TT acknowledges funding from EPSRC studentship 2275810.

References

1. Arya, S., and Mount, D. Approximate nearest neighbor searching. In *Proc. 4th Ann. ACMSIAM Symposium on Discrete Algorithms* (1993), pp. 271–280.
2. Bernstein, A., Burnaev, E., Sharaev, M., Kondrateva, E., and Kachan, O. Topological data analysis in computer vision. In *Twelfth International Conference on Machine Vision (ICMV 2019)* (2020), vol. 11433, SPIE, pp. 673–679.
3. Bubenik, P., and Dłotko, P. A persistence landscapes toolbox for topological statistics. *Journal of Symbolic Computation 78* (2017), 91–114.
4. Bubenik, P., et al. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res. 16*, 1 (2015), 77–102.
5. Bukkuri, A., Andor, N., and Darcy, I. K. Applications of topological data analysis in oncology. *Frontiers in artificial intelligence* (2021), 38.
6. Byrne, H. M., Harrington, H. A., Muschel, R., Reinert, G., Stolz-Pretzer, B., and Tillmann, U. Topology characterises tumour vasculature. *Mathematics Today* (2019).
7. Chatterjee, S., and Barcun, S. A nonparametric approach to credit screening. *J. Am. Stat. Assoc. 65*, 329 (1970), 150–154.
8. Detrano, R. Heart Disease Data Set. V.A. Medical Center, Long Beach and Cleveland Clinic. *UCI Machine Learning Repository* (1988).
9. Durand, D. Credit-rating formulae. *Risk Elements in Consumer Instalment Financing, Ch. 4* (1941).
10. Gidea, M., and Katz, Y. Topological data analysis of financial time series: Landscapes of crashes. *Physica A: Statistical Mechanics and its Applications 491* (2018), 820–834.
11. Henley, W. E. Statistical aspects of credit scoring. *Ph. D., Open University* (1995).
12. Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine 34*, 4 (2017), 43–59.
13. Mitic, P. A metric framework for quantifying data concentration. In *Proc. IDEAL 2019* (2019), Springer, pp. 181–190.
14. Ohlson, J. A. Financial ratios and the probabilistic prediction of bankruptcy. *Jnl. Accounting Research 18*, 1 (1980), 109–131.
15. Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., and Harrington, H. A. A roadmap for the computation of persistent homology. *EPJ Data Science 6* (2017), 1–38.
16. Oudot, S. Y. *Persistence theory: from quiver representations to data analysis*, vol. 209. American Mathematical Soc., 2017.
17. Riihimäki, H., Chachólski, W., Theorell, J., Hillert, J., and Ramanujam, R. A topological data analysis based classification method for multiple measurements. *BMC Bioinformatics 21*, 1 (2020), 1–18.
18. Wiginton, J. A note on the comparison of logit and discriminant models of consumer credit behavior. *Jnl. Fin. and Quant. Anal. 15*, 3 (1980), 757–770.
19. Wu, C., and Hargreaves, C. Topological machine learning for mixed numeric and categorical data. *Int. J. Artif. Intell. Tools 30* (2021), 1–18.