Question 1

A Comparison of Peer-to-Peer Loan Default Prediction Models

In this question we construct three different models for predicting defaults on a dataset of Peer-to-Peer loans: Logistic Regression, Support Vector Machines (SVM), and Decision Tree Classifiers. We perform 5-fold cross-validation for each model to ensure robustness and then compare their performance using traditional metrics.

This question is structured as follows: section one provides a brief introduction; section two discusses the data pre-processing, choices and assumptions made, and some of the details of the models themselves; section three presents the results; section four provides a discussion; and section five closes with concluding remarks and opportunities for further work.

1. Introduction

One can broadly define Peer-to-Peer (P2P) lending as the extension of credit to individuals or businesses through non-traditional means. By non-traditional we mean that, in general, lenders are not high-street financial institutions such as banks or building societies and most transactions occur online in a bid to reduce operating expenses.

One aspect remains central to the P2P industry though: Credit Risk. More specifically, companies are concerned about default risk. Reliable predictors for the likelihood of default are critical for businesses to survive. In this question we explored and evaluated several machine learning techniques for the prediction of defaults based on information typically provided during the application process.

2. Methodology

2.1 Data & Pre-processing

We were provided with a dataset containing loan information for the company Lending Club [1]. This held approximately 76,000 records with 21 features. To correctly apply our proposed models, we needed to pre-process the data.

Firstly, we considered which features to keep. Whilst all features could potentially have some bearing on whether an applicant is likely to default on their loan, we chose to take a marginally smaller subset of features to simplify our problem. The features we removed were: (*i*) Verification Status (*verification_status*), this generalised our results since the level and approach to verification may vary across the industry; (*ii*) Issue Date (*issue_d*), to simplify our problem and remove the time-series component. It would be interesting to explore whether certain time periods (e.g. pre-/post-financial crisis) give rise to structural changes in the likelihood of default, however, we do not consider that in this question. (*iii*) Earliest Credit Line on record (*earliest_credit_line*), as this information was grouped by year and was considered too general to be useful for our analysis; and (*iv*) number of Mortgage Credit Lines on record (*mort_acc*), as we felt that this information provided little additional value above that captured by the feature, Total Number of Credit Lines on record (*total_acc*). This left a total of 17 features.

We examined the number of missing values and found they existed in three features: Employment Length (*emp_length*); Revolving Utility of Credit Lines (*revol_util*); and Public Record Bankruptcies (*pub_rec_bankruptcies*). To rectify this, we used mean imputation for the first two features, however, we simply removed the missing values for the third feature as this was a highly unbalanced variable and mean imputation was inappropriate. In addition, it only accounted for a small fraction of the data (< 0.2%).

To prepare the inputs to the models, we standardised our numerical features. For the categorical features we used the One Hot Encoding [2] technique to transform the data. This involved expanding each categorical feature into its own binary indicator variable. An example of which is provided in Figure 1.

Application Type	**Application Type: *Individual***	**Application Type: *Joint***
Individual	1	0
Joint	0	1

Fig 1: An example of the transformation of categorical features using the One Hot Encoding technique. We can see we have moved from categorical to numerical data, at the expense of introducing new pseudo-features.

We deemed this approach superior to the simpler ordinal encoding technique as it does not introduce spurious ordering to our features which could disrupt the results of our models.

Finally, as we were working with predictive models, we needed to take care with the unbalanced nature of the dataset. In general, most people will not default on their loans, hence, we needed to account for this fact during the training of our models. We chose to undersample the majority class of the data to create a balanced dataset. This meant our final dataset, after pre-processing, contained ~27,500 records. Our feature of interest (*charged_off*) is binary: those that defaulted (indicated by 1) and those that did not (indicated by 0).

2.2 Logistic Regression [2]

We believed Logistic Regression was an appropriate method for consideration given the binary nature of our intended predictions: default versus no default. Logistic Regression relies on the Logistic Function,

$$f(\boldsymbol{\theta}) = \frac{1}{1+exp(-\boldsymbol{\beta}\cdot\boldsymbol{\theta})}.$$

This function (known as the Sigmoid function) produces a classification result in terms of probabilities, based on the n inputs from the feature space. Hence, the output is a value in the interval [0,1]. This means that we can assign an appropriate decision threshold, say 0.5, and classify all points lying above or below that threshold into their respective binary categories.

For our implementation we have used the Scikit-Learn package available in Python. As Question 2 focusses on regularization techniques, whereas the emphasis of this question is on comparison between models, we have removed the *LogisticRegression*(.) function's automatic regularization.

2.3 Support Vector Machines [3]

Support Vector Machines are a similar classification method and as such presented another appropriate choice for comparison. The SVM method seeks to find the optimal hyperplane that separates our binary output variable.

We can consider SVMs as an optimisation problem that seeks to maximise the distance (referred to as the margin) between the hyperplane and those output points that lie closest to the hyperplane (referred to as support vectors). We can extend SVMs to better account for non-separable data by introducing a soft-margin, that is accepting some points will inevitably be mis-classified. However, the true power of SVMs lies in its flexibility regarding the shape of the hyperplane. We can introduce a non-linear map (referred to as the kernel) which takes elements from our feature space, transforms them into a higher dimensional space, finds the optimal hyperplane there, then transforms back to our original space to improve the classification.

We have used the Scikit-Learn package available in Python for our implementation, during which we tested each of the four standard kernels provided. We found that the Gaussian Radial Basis Function performed best and, as this question is a comparison between different models, we only present these scores in our results section.

2.4 Decision Tree Classifier [2]

We also chose to consider Decision Tree Classifiers because they are easily interpretable and can be explained to non-technical audiences. They create a tree like structure starting from an initial or 'root' node before progressing through several layers. At each layer there are additional nodes which dictate a decision and split the data further eventually leading to classification.

Despite their simplicity Decision Tree Classifiers are susceptible to overfitting, that is, we can continuously split our data until we have an almost perfectly accurate tree. To address this problem, we performed a grid search on the maximum depth of the tree using the average AUC score of the test set as our indicator. We found that a maximum depth of 7 produced the best results and display the associated metrics in our results. Figure 2 shows the outcome of our grid search.

Once again, we have used the Scikit-Learn package available in Python, and we have specified the use of entropy (the information gain) as the parameter with which to assess our trees.

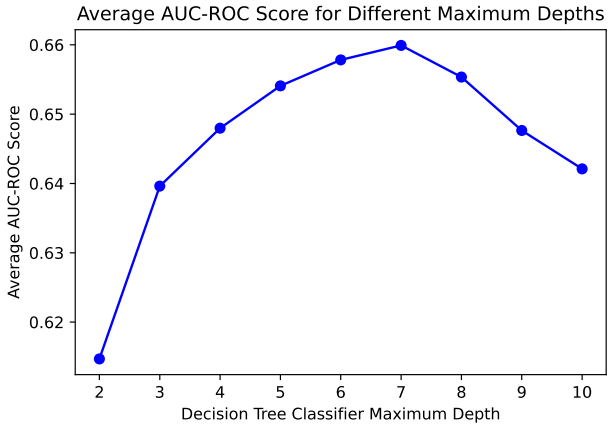

Fig 2: Graph showing the average AUC-ROC score for the grid search maximum depth of the cross-validated Decision Tree Classifier.

2.5 Cross-validation

To ensure robustness in our results we have used 5-fold cross-validation for all our models. This involves splitting the training and testing sets five different times, with each iteration using a different subset of the data (20% in the case of 5-fold cross-validation) as the test set.

3. Results

3.1 Interpretation of Results

The results of our models are displayed in Table 1. We present four commonly used metrics to evaluate our results: AUC-ROC Score; Accuracy; Precision; and Recall.

To interpret these metrics, we must first understand the confusion matrix, which shows a comparison of the actual outcome values to those predicted by our models. Figure 3 displays the definition of a confusion matrix. Whilst Figures 4–6 display the actual results for each model.

	Predicted: No Default	***Predicted***: Default
Actual: No Default	η_{00}	η_{01}
Actual: Default	η_{10}	η_{11}

Fig 3: The definition of a confusion matrix. We computed the confusion matrix for each model and calculated the metrics: Accuracy, Precision, and Recall. The results are displayed in Table 1.

The ROC curve plots a line with the False Positive Rate (FPR) on the x-axis against the True Positive Rate (TPR) on the y-axis. Where we define FPR and TPR as,

$$FPR = \frac{\eta_{01}}{\eta_{00} + \eta_{01}}$$

$$TPR = \frac{\eta_{11}}{\eta_{10} + \eta_{11}}.$$

If our model perfectly predicted all outcomes, we would see a vertical line traverse the y-axis, and a horizontal line extend from the point $y = 1$. In a purely random model, we expect the ROC curve to look like the line $y = x$. Using this ROC curve, we can compute the AUC Score which calculates the area under the ROC curve. Again, a perfect model would have an AUC score of 1, whilst a random model would have a score of 0.5. We used the random AUC score of 0.5 as a benchmark to evaluate the predictive power of our models.

We define the remaining three metrics as,

$$Accuracy = \frac{\eta_{00} + \eta_{11}}{\eta_{00} + \eta_{01} + \eta_{10} + \eta_{11}}$$

$$Precision = \frac{\eta_{11}}{\eta_{01} + \eta_{11}}$$

$$Recall = \frac{\eta_{11}}{\eta_{10} + \eta_{11}}.$$

We can interpret the accuracy as the number of datapoints our models correctly classified, that is, did our model predict defaults where a default truly occurred and likewise for no defaults. Precision can be considered as the proportion of those that truly defaulted versus the number our models predicted would default. Finally, we can interpret recall as those that we correctly predicted would default out of the total number that did indeed default. In all cases a score closer to 1 indicates theoretically better performance.

3.2 Presentation of Results

The first thing we notice is that our SVM performance was superior across all metrics. This was likely due to the non-linear (Radial Basis Function) kernel used in the SVM model, whereas Logistic Regression relied on a linear separation.

The AUC scores for all models were in the mid-to-high sixties, clearly outperforming our random model benchmark and indicating that each of them does indeed have some degree of predictive power. The scores for Logistic Regression and SVM were similar at approximately 69%, a 3% improvement over the simpler Decision Tree Classifier approach.

Similarly, the accuracy scores were well above our random model benchmark, with the SVM model producing an average score of 65.2%. This means that the model correctly classified the data into the two binary cases: default and no default an average of 65.2% of the time across the five test sets.

It is worth noting that we can be confident that this is a true reflection of the accuracy of the models because of the initial data pre-processing. If we had simply performed the analysis on the original unbalanced dataset our results could have been distorted by the significant presence of the majority class: no default.

The precision scores show a much narrower interval of values – all of which fall within a 1% range. This is an interesting outcome as it shows that a consistent number of cases are being predicted as default regardless of the model used. This can be seen by the similar numbers in the right-hand column of each of our confusion matrices (Figures 4–6).

Finally, our SVM model produced a recall score of 65.1%, meaning that of those cases that truly defaulted in the test set, our model correctly predicted 65.1% of them.

4. Discussion

Given the business context of the question, we propose the recall score is a more appropriate metric for evaluation than precision. In the P2P loan industry, default risk poses a significant concern. This means a model predicting no default when there was indeed a default is more problematic than predicting default in those that did not. The former could cause significant losses on issued loans, whereas the latter would simply reduce the number of loans issued in the first place.

This point, however, proves inconsequential in our analysis as the SVM comparatively outperforms across all metrics, thus we can infer that it is the best model of those considered.

Unfortunately, though, this superior relative performance does not imply good absolute performance. The results for all three models are disappointing. The best recall score of 65.1% still means that our model incorrectly classified over a third of cases, that is we predicted more than a third of cases would not default when they in fact did. One could certainly conclude that a business with such a loose approach to lending would not be solvent for long.

Model	AUC Score	Accuracy	Precision	Recall
Logistic Regression	68.9%	63.9%	64.2%	63.7%
Support Vector Machine	69.2%	65.2%	64.7%	65.1%
Decision Tree Classifier	66.0%	63.3%	63.7%	62.7%

Table 1: The performance metrics associated with each of our three proposed models: Logistic Regression, SVM, and Decision Tree Classifier. The results displayed are the averages obtained from 5-fold cross-validation.

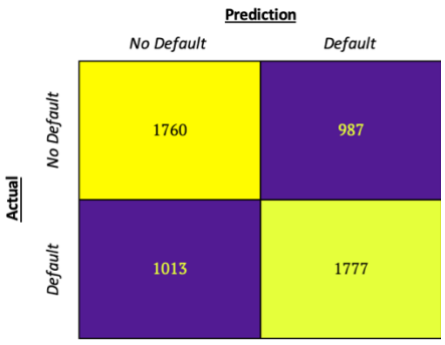

Fig 4: Logistic Regression Average Confusion Matrix

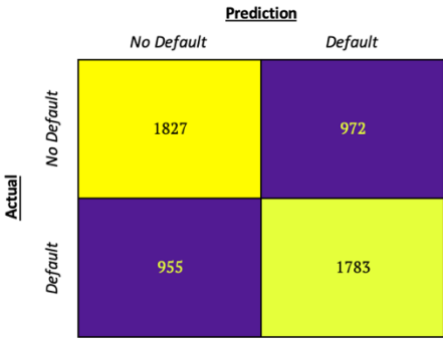

Fig 5: SVM Average Confusion Matrix

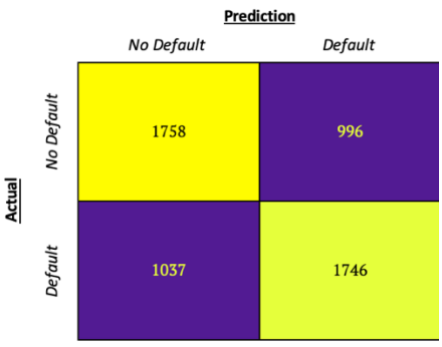

Fig 6: Decision Tree Average Confusion Matrix

There could be several reasons why these results are underwhelming. Firstly, as we might expect, loan data is inherently messy. Whilst the information collected during the application process and already in existence on an applicant's credit file will go some way towards indicating the likelihood of default, there will always be a stochastic element. This reflects the unique circumstances that every applicant will face and that cannot be captured during an application. For instance, a company may collapse during the term of the loan, leaving an employee that was previously unlikely to default unable to service their repayments. Of course, macro or economic factors are likely to be forecast elsewhere in the company and influence things such as their risk appetite.

Secondly, during our pre-processing we removed the time-series element of the data. We noted in Section 2 that this simplified our problem, but there are underlying consequences of this decision. By removing the time-series aspect we remove the ability to examine whether there is a credit environment regime change, by which we mean there could be two (or more) discrete periods within which predicting default could be substantially different. The dataset provides information for the range 2007–2017, hence, we could perhaps see different performances during the financial crisis between 2007–2010, and a relatively calmer period between 2011–2014.

Finally, the processed dataset used in our analysis is rather small at only 27,500 records. This allowed us to address the unbalanced nature and avoid misleading results but had the disadvantage of significantly reducing the volume available to train and test our models.

5. Conclusion

In this question we have compared the performance of three different methods for the prediction of P2P loan defaults. We can conclude that our models do indicate predictive power – with our SVM model being deemed superior – but sole reliance on such models would be inappropriate. In a business context though, as is so often the case with Machine Learning, the use of such models in conjunction with human oversight could bring additional efficiency, robustness, and a data-driven approach to loan applications.

We have also discussed some of the drawbacks that were encountered during our analysis, none of which are unsolvable. A natural extension of this work would be to develop some of these ideas further. For instance, one could use alternative techniques for the data pre-processing such as weighting the majority and minority classes to address the unbalanced nature of the data and use this to train and test the models. One could perform a deeper analysis of the Decision Tree Classifier model and consider whether the use of Forests improves performance. In addition, it would be interesting to perform detailed feature importance analysis to establish if there is a subset of features that drives predictions – this could then be utilised to improve the application process and remove unnecessary information capture. Alternatively, one could consider more advanced machine learning techniques such as Neural-Networks and perform further comparisons with our work.