ECS784P - Data Analytics
Assignment 1

Sarthak Agarwal - 220661227
MSc Computer Science

Introduction & Project Background

All industries, including real estate, security, bioinformatics, and the financial sector, are experiencing a revolution in process thanks to machine learning algorithms. One of the most time-consuming tasks in the banking business is the loan approval process. The efficiency, efficacy, and correctness of loan approval processes can be increased with the help of contemporary technology like machine learning models. In order to forecast loan eligibility, this study provides three machine learning algorithms: Support Vector Machine (SVM), K-Nearest Neighbour (KNN), and Logistic Regression (LR). The historical dataset "Loan Eligibility Dataset," which is accessible on Kaggle and licenced under Database Contents License (DbCL) v1.0, was used to train the models. The research's findings demonstrated great performance accuracy, with the SVM algorithm scoring highest and KNN scoring lowest. In terms of precision-recall and accuracy, the Models surpassed two of the three loan prediction models that were discovered in the literature.

The banking industry, like many other businesses, is increasingly striving to take use of the opportunities provided by contemporary technologies to enhance their operations, boost productivity, and cut costs. Predictive analytics was the most widely used machine learning feature for applications in the global banking sector in 2020, claims. The capacity of most lending platforms to assess credit risk will determine whether they are successful or not. Any financial organisation that approves loans faces a difficult task. Before extending credit to debtors, the bank determines if they are good or bad (Paidoff) (Collection). In order to speed up the decision-making process and determine whether an application receives a loan or not, this article focused on using Machine Learning (ML) models to forecast loan eligibility. In this work, the goals are to (1) prepare the data for modelling by cleaning and pre-processing it, (2) do exploratory data analysis (EDA) on the dataset, (3) develop several machine learning models to predict loan eligibility, and (4) assess and compare the various models developed.

Literature Review

With the purpose of finding and contrasting the best-fit ML-based models for credit risk assessment, the authors of [1] conducted a comprehensive literature review. The authors' goal was to demonstrate the various ML algorithms used by researchers to evaluate rural borrowers' credit, particularly those with a limited loan history. Their findings demonstrated that the machine learning (ML) techniques we used in this study were widely used and produced excellent outcomes.

Banks are searching for more efficient ways to conduct the loan approval process due to the negative effects that low loan payback rates have on them on a global scale. Using Random Forest, XGBoost, GBM, and neural network machine learning models, the authors of [2] assessed the loan default prediction of the Chinese peer-to-peer (P2P) market. Their four models, with RF being the best, were more accurate than 90% of the time. This study is highly related in terms of the techniques and algorithms employed, their goal was similar to us which is to forecast P2P loan default.

To forecast loan acceptance using bank direct marketing data, [3] used a variety of ensemble ML approaches, including AdaBoost, LogitBoost, Bagging, and Random Forest model. AdaBoost had the highest accuracy, according to their research, at 83.97%. Our study differed significantly from previous research in that we move ahead with imbalanced dataset as we are only concerned about loan being Paidoff, and as a result, models performed better.

In order to forecast consumers' creditworthiness and assist banks in

developing an automated risk assessment system, the research by [4] examined actual bank credit data. They used a variety of machine learning (ML) algorithms, including decision trees, neural networks, naive Bayes, KNN, and ensemble learning algorithms. Their models' accuracy varied between 80% and 76%, which is similarly less than that of our models.

Furthermore to implement ML models and to calculate f1, precision & recall for our three models [5] had the same model implementation and helped in setting us classification report metrics.

Data Processing

The historical dataset "Loan Eligibility Dataset," which is accessible on Kaggle and licenced under Database Contents License (DbCL) v1.0, was used to train the models. This dataset is about past loans. The data set includes details of 346 customers whose loan are already paid off or defaulted. Description of dataset is shown in Fig 1

Field	Description
Loan_status	Loan is paid off on in collection
Principal	Principal loan amount
Terms	Payment Cycle
Effective_date	Loan start date
Due_date	Loan repayment date
Age	Age of borrower
Education	Education of borrower
Gender	Gender of borrower.

Fig 1 : Description of Dataset

We have two date columns, effective_date & due_date, original date format is DD/MM/YYYY, we will convert it to YYYY-MM-DD which is easier to parse as a parameter to ML model. In column loan_status we see that out of 346 rows, 260 borrowers repaid (paidoff) their loan whereas 86 were defaulters (collection). This will be our target column i.e. what we need to predict. It has two classes COLLECTION & PAIDOFF, which indicates the problem is to solve binary classification, since we are focused on identifying the good borrowers i.e. paidoff, our evaluation metric will focus on paidoff accuracy.

The correlation of data is explored in five ways. First, Gender and Principal amount, as shown in Fig 2.

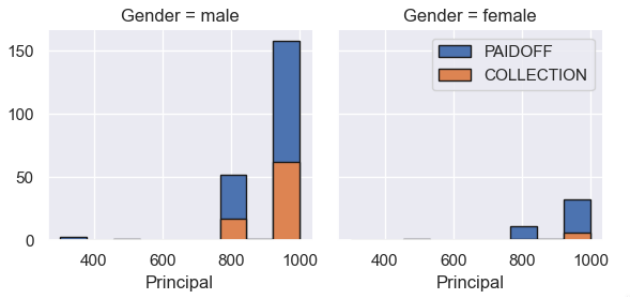

Fig 2 : Correlation of Gender & Principal Amount

Majority of borrowers have principal amount greater than 600 & there are more male borrowers than female.

Second, Gender and Age, as shown in Fig 3.

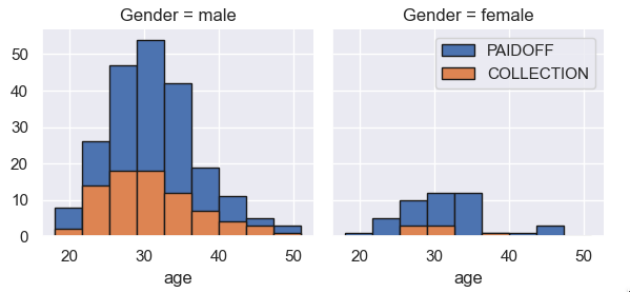

Fig 3 : Correlation of Gender & Age

Male tend to borrow more between the age of 20 & 40, which might indicate home/ education loan, whereas female has less age window for borrowing which is between the age of 25 & 35, indicating that female take loan more strategically that's why the compressed age bar.

Third, feature selection is performed by dropping column due_date as we already have info of start_date & terms which ultimately leads to due_date, furthermore the date in itself is distributed parameter across a calendar, better to consider the day of the

week, which is extracted from date, as shown in Fig 4.

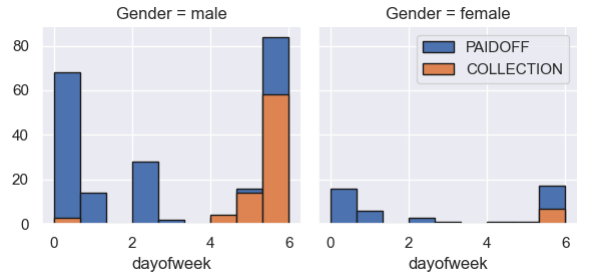

Fig 4 : Correlation of Gender to Loan Start Day

We see that borrowers who receive loans at the end of the week do not repay them, so let's utilise feature binarization to establish a threshold value lower than day 4, which means all day > 3 will be represented as 1 & day < 4 will be 0.

Four, Gender & Loan Status, i.e. percentage of male & female repaying their loan, as shown in Fig 5.

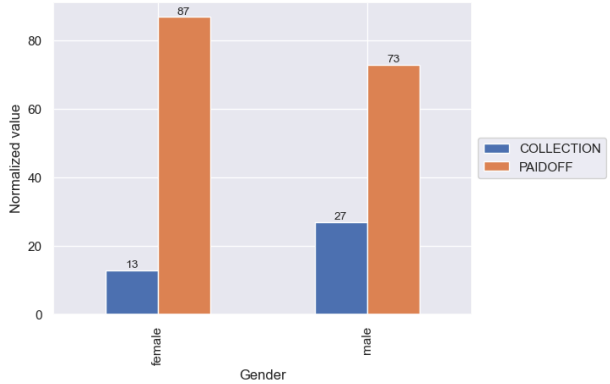

Fig 5 : Percentage of Paidoff & Collection, based on Gender

87 % of female pay their loans while only 73 % of males pay their loan. We will use One-Hot Encoding technique to convert male to 0 & female to 1.

Five, we see Education & Loan Status correlation, as shown in Fig 6.

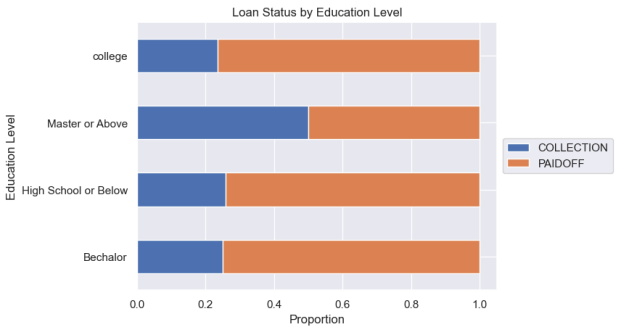

Fig 6 : Loan Repayment based on Education

We see that for Master or Above, the probability of Collection is higher, since we are only looking for paidoff we will remove Master or Above category and then One-Hot Encode the column education.

Normalised data is used to remove bias from data, for example Principal value 1000 & 100 would make weightage of principal differ from row to row, to handle this issue we use Standard Scaler, to created standardized features by removing the mean and scaling to unit variance.

Learning Methods

When it comes to binary classification problems such as loan repayment, there are several machine learning algorithms that can be used, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Logistic Regression. Here's a justification for using each of these algorithms:

K-Nearest Neighbors (KNN): KNN is a simple and intuitive algorithm that can be used for binary classification. It works by finding the k-nearest data points to a given point and classifying it based on the majority class of those neighbours. In the context of loan repayment, KNN can be used to predict whether a borrower is likely to repay their loan or not based on the characteristics of their loan application and their past credit history.

Support Vector Machines (SVM): SVM is a powerful algorithm that can be used for binary classification. It works by finding the hyperplane that maximizes the margin between the two classes. In the context of loan repayment, SVM can be used to predict

whether a borrower is likely to repay their loan or not based on the characteristics of their loan application and their past credit history. SVM is particularly useful when the data is non-linearly separable, meaning that the classes cannot be separated by a simple linear boundary.

Logistic Regression: Logistic Regression is a widely used algorithm for binary classification. It works by modelling the probability of the positive class as a function of the input variables. In the context of loan repayment, logistic regression can be used to predict the probability of a borrower repaying their loan based on the characteristics of their loan application and their past credit history. It's also a simple and interpretable algorithm, making it easy to understand and explain the predictions to stakeholders.

In conclusion, KNN, SVM, and Logistic Regression are all suitable algorithms for loan repayment binary classification. The choice of algorithm will depend on the specific characteristics of the data and the trade-offs between accuracy, interpretability, and computational efficiency.

Analysis, Testing, Results

KNN

To demonstrate the impact of the k parameter on the accuracy of the KNN algorithm, manual traversal is a valid approach. This involves training and testing the KNN algorithm on the dataset multiple times, each time with a different value of k, and recording the resulting accuracy. By doing this, you can create a plot or graph that shows how the accuracy of the KNN algorithm changes as the value of k increases or decreases. This can be a useful visualization to help understand the behaviour of the KNN algorithm on the given dataset and to choose an appropriate value of k, as shown in Fig 7.

Fig 7 : Accuracy for different neighbours in KNN

K = 3 is the best, so we choose 3 as neighbours in KNN

SVM

Choose the best kernel for SVM, we test four of them, i.e. linear, poly, rbf & sigmoid.

For binary classification issues like loan repayment, it is frequently observed that the "rbf" and "poly" kernel functions in SVM provide superior accuracy than the "linear" and "sigmoid" kernel functions. This is so because, in contrast to the linear and sigmoid kernels, the "rbf" and "poly" kernels may reflect more complex decision boundaries.

The "linear" kernel presumes that a straight line may be used to divide the data, which may not be the case for many real-world datasets. Although the "sigmoid" kernel can simulate non-linear decision boundaries, in actual use it frequently performs less well than the "rbf" and "poly" kernels.

The "rbf" kernel is a popular choice for SVM as it can model non-linear decision boundaries by translating the input characteristics into a higher-dimensional space. Non-linear decision boundaries can also be modelled using the "poly" kernel, however this time, polynomial functions rather than radial basis functions are employed.

In our case poly & rbf performed the best with same accuracy as shown in Fig 8.

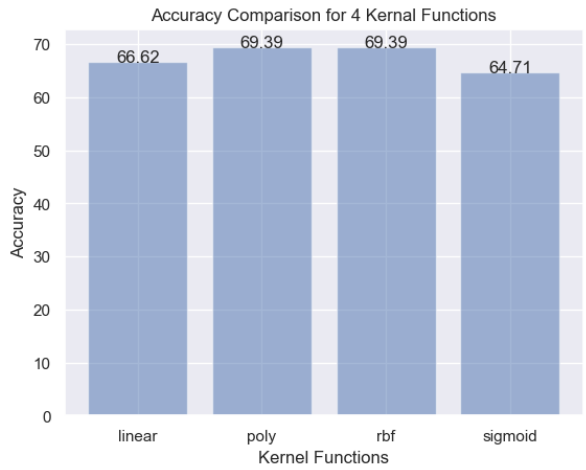

Fig 8 : SVM Kernel Accuracy

For our model we will use rbf kernel in SVM.

Logistic Regression

In logistic regression, the cost function optimization algorithm is referred to as the solver parameter. The solvers "lbfgs," "liblinear," "newton-cg," "sag," and "saga" are the most often utilised. For small and medium-sized datasets, "lbfgs" and "newton-cg" solvers are typically advised, but "sag" and "saga" solvers are typically advised for large datasets. Generally speaking, when the dataset is sparse just like ours, the "liblinear" solver is favoured.

A hyperparameter that regulates the regularisation strength in logistic regression is called the C value. The regularisation strength will increase with a low value of C, reducing overfitting of the model. On the other side, a high value of C will result in a weaker regularisation, which can help the model perform more accurately on the training set of data. Via hyperparameter tuning, the ideal value of C should be ascertained, as shown in Fig 9.

```
Parameter : 1, Accuracy for C = 0.1, solver = newton-cg : 50.25
Parameter : 2, Accuracy for C = 0.1, solver = lbfgs : 50.25
Parameter : 3, Accuracy for C = 0.1, solver = liblinear : 51.74
Parameter : 4, Accuracy for C = 0.1, solver = sag : 50.25
Parameter : 5, Accuracy for C = 0.1, solver = saga : 50.25

Parameter : 6, Accuracy for C = 0.01, solver = newton-cg : 49.57
Parameter : 7, Accuracy for C = 0.01, solver = lbfgs : 49.57
Parameter : 8, Accuracy for C = 0.01, solver = liblinear : 58.54
Parameter : 9, Accuracy for C = 0.01, solver = sag : 49.57
Parameter : 10, Accuracy for C = 0.01, solver = saga : 49.57

Parameter : 11, Accuracy for C = 0.001, solver = newton-cg : 51.8
Parameter : 12, Accuracy for C = 0.001, solver = lbfgs : 51.8
Parameter : 13, Accuracy for C = 0.001, solver = liblinear : 67.31
Parameter : 14, Accuracy for C = 0.001, solver = sag : 51.8
Parameter : 15, Accuracy for C = 0.001, solver = saga : 51.8
```

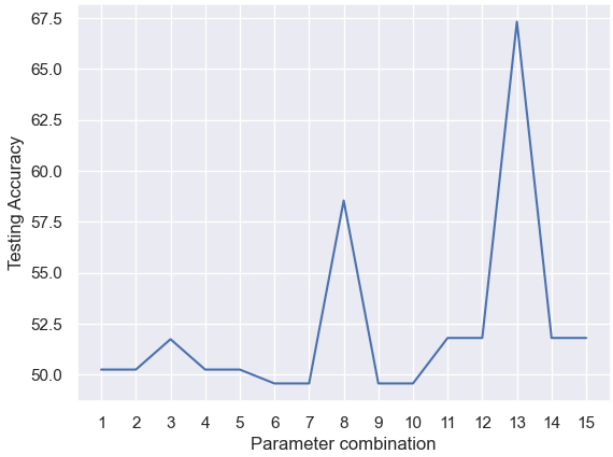

Fig 9 : Hyper-parameter Tuning for Logistic Regression

Parameter 13 had the highest accuracy which was c = 0.001 & solver = liblinear.

Result

To test our model we use a test dataset, and we perform the same pre-processing and feature selection that we performed on train set, after that we will predict the loan status using KNN, SVM, & LR. To calculate error/ accuracy we use scoring systems such as Jaccard, F1, Precision, Recall. Scores represented as heatmap is shown in Fig 10.

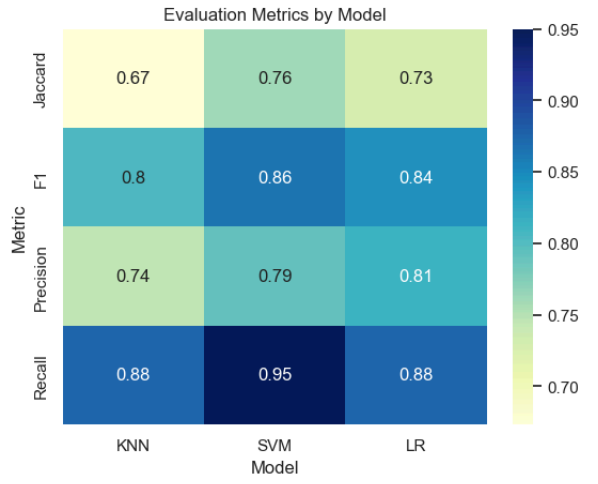

Fig 10 : Heatmap of Scores

Visualizing the heatmap in form of bar chart, as shown in Fig 11.

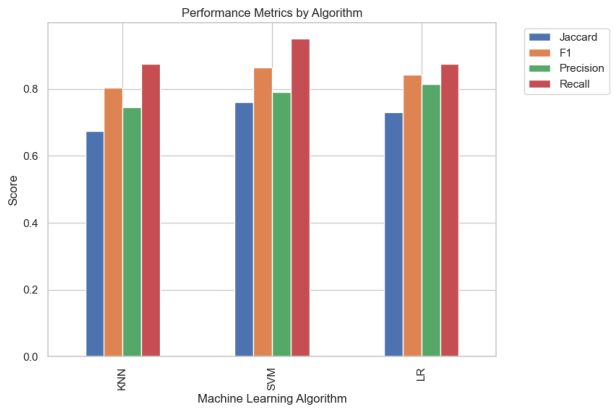

Fig 11 : Bar Chart of Scores

Group all scores to get an overview on performance of models.

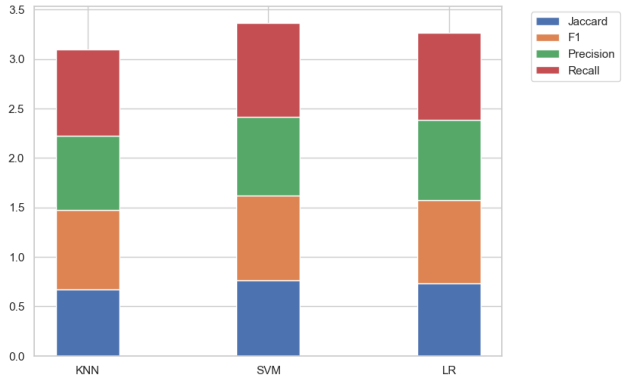

Fig 12 : Stacked Score for Classification Model

To understand in-depth performance of a model we need to look at each score individually and understand what do they imply.

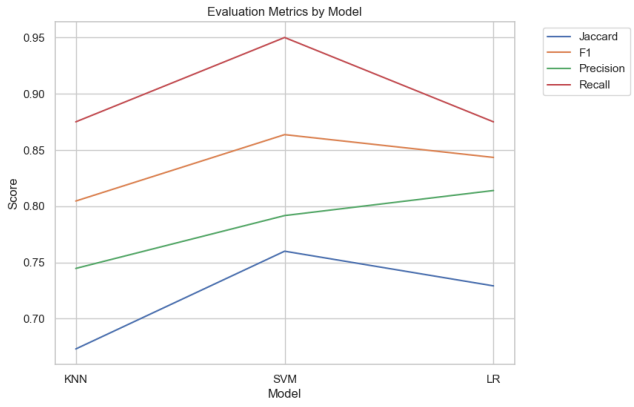

Fig 13 : Trend of Individual Scores in Classification

Looking at the Jaccard index, which measures the similarity between the predicted labels and the true labels, SVM has the highest score of 0.76, followed by Logistic Regression with a score of 0.73, and then KNN with a score of 0.67. This suggests that SVM and Logistic Regression are better at predicting the loan repayment status compared to KNN.

In terms of the F1 score, which is the harmonic mean of precision and recall, SVM has the highest score of 0.86, followed by Logistic Regression with a score of 0.84, and then KNN with a score of 0.8. This suggests that SVM and Logistic Regression have better balance between precision and recall compared to KNN.

Looking at precision, which measures the proportion of true positives among all predicted positives, Logistic Regression has the highest score of 0.81, followed by KNN with a score of 0.74, and then SVM with a score of 0.79. This suggests that Logistic Regression is better at identifying true positives among all predicted positives.

In terms of recall, which measures the proportion of true positives among all actual positives, SVM has the highest score of 0.95, followed by KNN with a score of 0.88, and then Logistic Regression with a score of 0.88. This suggests that SVM is better at identifying all actual positives.

Overall, SVM appears to be the best model for this particular problem based on the evaluation metrics.

Conclusion

The problem of binary classification for loan repayment has been addressed using three popular machine learning algorithms: KNN, SVM, and Logistic Regression.

The models were evaluated using several common evaluation metrics such as Jaccard, F1, Precision, and Recall. Based on these metrics, it appears that SVM has the highest overall performance, with higher values of Jaccard, F1, Precision, and Recall compared to KNN and Logistic Regression.

However, it's important to note that the choice of algorithm may also depend on other factors such as the size and complexity of the dataset, as well as the specific requirements of the problem. Additionally, it may be beneficial to

perform hyperparameter tuning for each algorithm to obtain the best possible performance. Some possible limitations of these models include:

Data imbalance: The dataset may be imbalanced, with a large number of examples belonging to one class (e.g., loan repayment) and relatively few examples belonging to the other class (e.g., loan default). This can lead to biased models and poor performance on the minority class.

Missing data: The dataset may have missing values or incomplete information, which can affect the performance of the models.

Non-linearity: The relationship between the input features and the target variable (i.e., loan repayment) may be non-linear, which can limit the performance of linear models such as Logistic Regression.

Future improvements or directions for this problem include:

Feature engineering: It may be beneficial to perform feature engineering to extract more meaningful features from the dataset, or to transform the existing features to make them more relevant for the prediction task.

Ensemble methods: Ensemble methods such as Random Forest or Gradient Boosting can be used to combine multiple models and improve the overall performance.

Neural networks: Neural networks such as Multilayer Perceptron or Convolutional Neural Networks (CNNs) can be used to model complex non-linear relationships between the input features and the target variable.

Explainability: It may be beneficial to improve the explainability of the models, so that the decisions made by the models can be easily understood and validated by domain experts. This can be particularly important in domains such as finance, where transparency and accountability are crucial.

References

[1] A. Kumar, S. Sharma, & M. Mahdavi, "Machine Learning (ML) Technologies for Digital Credit Scoring in Rural Finance: A Literature Review." Risks 9.11 (2021): 192

[2] J. Xu, Z. Lu, and Y. Xie, “Loan default prediction of Chinese P2P market: a machine learning methodology.” Scientific Reports, 2021, Vol. 11(1), pp. 1- 19.

[3] H. Meshref, "Predicting Loan Approval of Bank Direct Marketing Data Using Ensemble Machine Learning Algorithms."International Journal of circuits, systems, and signal processing. 2020, Vol. 14, pp. 914-922 DOI: 10.46300/9106.2020.14.117

[4] A.S. Aphale, and S.R. Shinde, “Predict Loan Approval in Banking System Machine Learning Approach for Cooperative Banks Loan Approval.” International Journal of Engineering Research & Technology (IJERT). 2020, Vol. 9 pp. 991-995

[5] D. J. P. Reddy, M. Gunasekaran and K. K. S. Sundari, "An Effective Approach for the Prediction of Car Loan Default Based-on Accuracy, Precision, Recall Using Extreme Logistic Regression Algorithm and K-Nearest Neighbours Algorithm on Financial Institution Loan Dataset," 2022 International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates, 2022, pp. 1-5, doi: 10.1109/ICCR56254.2022.9995969.