RESEARCH METHODS AND PROFESSIONAL PRACTICE

Literature review

Ruth Allison
ra22895

Can machine learning accurately classify customers based on their demographic and behaviour data?

Introduction

The ability to predict customer behaviour is necessary for businesses to compete in a global marketplace. Being able to accurately classify customers into 'likely to buy' and 'not likely to buy' can allow targeted marketing campaigns to maximise opportunities and minimise overly expensive or low yield activities.

The ability to classify customers allows more effective use of marketing resources. As Javadpour et al. (2021) discuss, banks need to automate the identification of target customers as numbers increase, as manual classification becomes impractical due to the ever-increasing volume of data requiring processing.

Customer data can be broadly classified into demographic and behavioural. Demographic data typically includes factors such as age, gender, marital status, level of education, and employment details. Behavioural variables may include frequency of purchases, average spend and tendency to buy specific items. Due to the large number of both factors and customers in a dataset it is impossible to determine relationships manually however machine learning can be applied to this problem.

This literature review will look at a range of algorithms for classifying customers. It will consider the availability of data, data pre-processing methods, feature selection methods, model selection and performance metrics and attempt to draw meaningful conclusions from the findings

Review of existing literature

Data collection

Whilst reviewing the literature for this paper it was seen that there are two main sources for datasets. Half of the papers reviewed used existing datasets available from the UC Irvine Machine Learning Repository (UCI) whilst the other half used real-world datasets. According to Brownlee (2019) the advantage of using pre-existing UCI datasets is they are well studied so expected 'good' results are known; however, the datasets are already cleaned so potentially useful data may have been removed. In comparison real-world datasets require much more effort and time to pre-process as there may be missing values, duplicates, and noisy data to address along with normalisation or standardisation, and data reduction. Only once this is complete can the dataset be classified. Whilst the UCI datasets are convenient to use for comparing classifiers, they may not be suitable for answering questions that require specific features not present in the pre-prepared datasets. In these cases, a real-world dataset must be obtained.

For real-world datasets Siva Subramanian et al. (2023) recommend the removal of noisy data and the use of mean methods for missing numerical values. However, Chen et al., (2021) replace missing values in categorical information with zeros. According to Firdose (2023), replacing a missing value with the mean for that feature should be treated with caution as it can skew the distribution. It is also only appropriate to use the mean if the dataset is normally distributed. If the data has outliers or is very skewed, then the median value would be more appropriate to avoid further skewing the dataset.

The decision on whether to use feature scaling on the dataset is dependent on the shape of the variables, and algorithms being used. Bhandari (2023) states that "Feature scaling…transforms the values…to a similar scale…to ensure all features contribute equally … and to avoid the domination of features with larger values". Following the removal of noise, duplicates, and missing values Chaubey et al. (2022) uses standard scaling whereas Ozan, (2018) uses normalisation to pre-process their dataset. As further explained by Bhandari (2023) standardisation arranges data around the mean and with a unit standard deviation which changes the shape of the distribution and is less sensitive to outliers. Normalisation scales the data, typically to values between 0 and 1, keeping the original distribution but it is more sensitive to outliers. Both methods allow features with widely varying scales to be more easily compared.

Another key issue highlighted by Chen et al. (2021) is imbalanced data, where there is a large disparity in the ratio between positive and negative outcomes. This can lead to poor classification performance, particularly when the positive to negative ratio is lower than 1:3. According to Brownlee (2021b), commonly used methods to address this are undersampling and oversampling. Undersampling deletes examples from the larger class which may lead to loss of data. Oversampling duplicates examples from the smaller class and may lead to overfitting with some models.

An additional consideration when collecting customer data is the uncertainty around which features are going to be most relevant for accurate customer classification. It is likely that data useful for predicting whether a bank customer is likely to buy a mortgage is different than for identifying customer preferences for supermarket own brand verses Heinz baked beans. A limitation of the data collected, as highlighted by Chen et al. (2021) is that datasets may not be complete due to customer privacy laws. Some information may not be available at all, or its use may be heavily restricted as customers may not have given permission for it to be used. This may be addressed by including the uses that the data will be put to in a customer agreement that must be signed.

Feature selection

The data available for analysis may contain demographic, environmental, or behavioural information. Not all features represented in the dataset will impact on the target variable so these should be discarded to avoid adding unnecessary processing requirements during the model training phase. Identifying the most important features for analysis will ensure maximum impact for training the chosen model. A common method, as used by Chaubey et al. (2022) and Siva Subramanian et al. (2023) is to use a chi-squared test to rank the features.

The optimum number of features to use is generally obtained through a process of testing the algorithm with different combinations of features. As Chaubey et al. (2022) observe "The accuracy of the model does not depend on the feature selection (e.g., the best features that you have selected may or may not give the best accuracy)". This may be because of unknown correlations between variables. Kuhn & Johnson (2018) explain that most feature selection techniques are univariate, evaluating each predictor on its own. The existence of correlated predictors makes it possible to select predictors that appear important but are not actually useful. This necessitates a trial-and-error approach to optimal feature selection.

Classifications methods

Computer classification models are a relatively new development with huge steps forwards taken in just a few years. Initial models, such as those posited by Chiu (2002), used case-based reasoning to

determine the likelihood of a customer purchasing insurance. This was based on comparing a new customer (case) to existing cases and determining the similarity however they stated that "Most of the time weighting values are determined using human judgement and thereby the retrieved solution(s) cannot always be guaranteed".

By 2015, according to Das (2015) the most commonly used classifiers were Naïve Bayes, K nearest neighbour (KNN) and Support Vector Machines (SVM) with additional methods, including ensemble methods, being developed rapidly. As explained by Brownlee (2021a) ensemble techniques fall into three main classes – bagging, boosting, and stacking. Bagging uses samples of the same dataset with decision trees and averaging the result. Boosting is the sequential addition of methods and stacking uses a base level with two classifiers that are trained on the same dataset, followed by a meta level that uses a KNN classifier trained on the outputs of the base level.

Further progress is demonstrated by Chaubey et al. (2022) who compared thirteen methods of classification, including ensemble techniques as well as a dummy classifier. Although a dummy classifier is not used for predicting classes it provides a useful baseline for comparison with other models. Alongside the more traditional methods the ensemble techniques that they compared were random forest (bagging); AdaBoost and XgBoost (boosting); and SvmAda, RfAda and KnnSgd (stacking). It is likely that these ensemble methods will continue to develop in the future as datasets continue to expand in size and businesses struggle to gain new and retain existing customers in an increasingly data driven world.

Measuring model performance

There are several metrics for determining the performance of a classifier, however all start by determining the number of true and false positive and negative results. Which metrics are chosen depends on the purpose of the analysis and, as mentioned by Mishra (2018) just using accuracy is not a true measure of a model.

Common metrics that are used:

- *Accuracy* is the proportion of correctly classified instances (both true positive and true negative) and should be close to 1 (normally quoted as a percentage). According to Harikrishnan (2019) caution should be used with this metric when imbalanced datasets are used, however. For example, a sample of 100 people has 90 who are healthy and 10 who are ill. If everyone is classified as healthy then the accuracy of the model would be 90% however all the people who are ill have been wrongly classified.
- *Precision* is the proportion of all identified positives that are true positives. For a good classifier the precision should be 1 meaning there are no false positives.
- *Recall* (also called sensitivity or the true positive rate) should also be 1 for a good classifier meaning there are no false negatives.
- Specificity (true negative rate) is inversely proportional to sensitivity.
- *F1 score* takes both precision and recall into account, so is a better measure than accuracy. The F1 score is 1 when both precision and recall are 1 (Harikrishnan, 2019)
- *Receiver operating characteristics (ROC)* curves are a plot of the true positive rate (y-axis) against the false positive rate (x-axis) at specified intervals. According to Narkhede (2018) the area under a ROC curve is determined, with a good model having an area close to 1, showing it is able to distinguish between positive and negative classes very accurately.

Analysis of the literature found multiple classifiers being compared, including single algorithms and ensemble methods. Methods popular in 2015, such as Naïve Bayes, KNN and SVM are still being used by researchers today, as presented in the following.

The Naïve Bayes model was used in half of the papers reviewed with Rusli et al. (2023) finding it 67.7% accurate for predicting the use of coupons sent to customers, whereas Das (2015) found an accuracy of 95% for predicting if existing bank customers would be interested in further products. Das (2015) also had a recall value of 98% and a specificity value of 76% showing the model performed well in correctly identifying positive values but less well for negative values. In this case the ability to correctly identify customers likely to purchase products probably outweighs the customers who received information about products that they were not interested in. Chaubey et al. (2022), Rajak et al. (2022) and Siva Subramanian et al. (2023) recorded accuracies between 82% and 89%. As Rajak et al. (2022) did not use any other evaluation metrics it is difficult to say if the model was suitable for predicting bank customer behaviour.

The SVM model was also used in several papers, with accuracies of 80%, 88%, 89% and 99% reported. Although the value of 99% accuracy seems very good Javadpour et al. (2021) did not use any other metrics to evaluate their model. In fact, Javadpour et al. (2021) used six models and obtained an accuracy of between 97% and 99% for each with no other measures; thus, a meaningful comparison of the effectiveness of these models for determining customer credit ratings is not possible without additional testing. Chaubey et al. (2022) reported an accuracy of 89% and a ROC curve area of 0.97 showing the SVM model was effective at distinguishing between positive and negative values.

Chaubey et al. (2022) compared several ensemble methods, finding an accuracy of 90% for the Random Forest method (bagging); 90% for Adaboost and 91.7% for Xgboost (boosting); and 89% for SvmAda, 91% for RfAda and 92.4% for KnnSgd (stacking). The ROC curve area for the KnnSgd model was 0.96 so this model seems to be very effective at determining customer purchasing behaviour. The results from Chaubey et al. (2022) are compared using ROC curves which made a visual comparison between the many models possible as the curves are more intuitive to compare than a large table of figures.

Conclusion

In terms of overall accuracy there does not appear to be any consensus on which is the 'best' algorithm to use. In the papers reviewed, the most accurate classification method is different for each, with accuracy levels depending on the criteria used.

Most papers reviewed looked at financial services customers, so it is not clear if these results are transferrable to other sectors. It would be useful to establish if there are optimum methods depending on the customer or product type. These could be evaluated by the degree of accuracy acceptable in the prediction model. For example, in the classification of cancer patients it may be preferable to have false positives than false negatives so patients who have cancer do not miss out on life saving treatment, whereas a spam detector would be better with false negatives rather than false positives to avoid important emails being blocked.

In conclusion the literature shows that customer classification can be achieved by careful use of feature selection methods and appropriate algorithms to a degree of accuracy that may be appropriate for the business needs analysed.

References

Bhandari, A. (2023) Feature Engineering: Scaling, Normalization, and Standardization. Available from: https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/ [Accessed 30 December 2023]

Brownlee, J (2019) Practice Machine Learning with Datasets from the UCI Machine Learning Repository. Available from: https://machinelearningmastery.com/practice-machine-learning-with-small-in-memory-datasets-from-the-uci-machine-learning-repository/ [Accessed 30 December 2023]

Brownlee, J (2021a) A gentle introduction to ensemble learning algorithms. Available from: https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/#:~:text=The%20three%20main%20classes%20of,on%20your%20predictive%20modeling%20project. [Accessed 30 December 2023]

Brownlee, J (2021b) Random oversampling and undersampling for imbalanced classification. Available from: https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/ [Accessed 30 December 2023]

Chaubey, G. et al. (2022) Customer purchasing behavior prediction using machine learning classification techniques. *Journal of Ambient Intelligence and Humanized Computing [Preprint].* DOI: https://doi.org/10.1007/s12652-022-03837-6

Chen, S. et al. (2021) Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine. *Expert Systems with Applications* 173(2): 114756. DOI: https://doi.org/10.1016/j.eswa.2021.114756

Chiu, C. (2002) A case-based customer classification approach for direct marketing. *Expert Systems with Applications* 22(2): 163-168. DOI: http://dx.doi.org/10.1016/S0957-4174(01)00052-5

Das, T. (2015) 'A customer classification prediction model based on machine learning techniques', 2015 Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) 321-326. DOI: https://doi.org/10.1109/ICATCCT.2015.7456903

Firdose, T. (2023) Filling missing values with Mean and Median. Available from: https://tahera-firdose.medium.com/filling-missing-values-with-mean-and-median-76635d55c1bc [Accessed 30 December 2023]

Harikrishan, N.B. (2019) Confusion matrix, accuracy, precision, recall, F1 score. Available from: https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd [Accessed 30 December 2023]

Javadpour, A. et al. (2021) Improving the Efficiency of Customer's Credit Rating with Machine Learning in Big Data Cloud Computing. *Wireless Personal Communications* 121(4): 2699–2718. DOI: https://doi.org/10.1007/s11277-021-08844-y

Katyayan, A., et al. (2022) 'Analysis of Unsupervised Machine Learning Techniques for Customer Segmentation'. In: Chen, J.IZ., Wang, H., Du, KL., Suma, V. (eds) *Machine Learning and Autonomous Systems. Smart Innovation, Systems and Technologies,* vol 269. Springer, Singapore. DOI: https://doi.org/10.1007/978-981-16-7996-4_35

Kuhn, M. & Johnson, K. (2013) *Applied Predictive Modelling*. 1st ed. Springer

Mishra, A. (2018) Metrics to evaluate your Machine Learning Algorithm. Available from: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234 [Accessed 30 December 2023]

Narkhede, S. (2018) Understanding AUC-ROC curve. Available from: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5 [Accessed 30 December 2023]

Ozan, S, (2018) 'A Case Study on Customer Segmentation by using Machine Learning Methods', 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 28-30 September. 1-6, DOI: https://doi.org/10.1109/IDAP.2018.8620892

Rajak, A. et al. (2022) 'Learning Paradigms for Analysis of Bank Customer', *Proceedings of the Third International Conference on Sustainable Computing*, pp. 115–124. DOI: https://doi.org/10.1007/978-981-16-4538-9_12

Rusli, N. et al. (2023) 'A Comparative Study of Machine Learning Classification Models on Customer Behavior Data'. *Communications in Computer and Information Science*. 222–231. DOI: https://doi.org/10.1007/978-981-99-0405-1_16

Siva Subramanian, R. et al. (2023) ‘Enhancing Customer Prediction Using Machine Learning with Feature Selection Approaches’. *Lecture Notes in Networks and Systems*. 45–57. DOI: https://doi.org/10.1007/978-981-19-7402-1_4