Profusion Technical Assessment
Credit Default Prediction Algorithm in Bank Marketing Dataset

Daniel Elston
Queen Mary University of London
MSc Data Science and Artificial Intelligence

March 7, 2022

Abstract

An artificial intelligence (AI) algorithm has been developed for the purpose of predicting whether a bank customer will default on a credit card. A test accuracy of 79% was achieved using an Extreme Gradient Boost (XGB) classifier. A recall of 96% is a highly attractive metric for the use of this algorithm, only misclassifying 4% of the identified potential defaulters. A large amount of algorithmic bias was determined, which, if removed, could produce a more accurate algorithm. This project further gives an excellent example of how algorithmic bias is effecting machine learning algorithms and how it can be mitigated with correct data gathering methods.

The raw data was initially used to predict if a customer will make a term deposit from telemarketing statistics. Approximately 25% of the default label were initially labelled 'unknown'. For the use of this analysis, 'unknown' was changed to 'yes'. The rest of the default feature retained their original values, 'no'. This dataset is therefore theoretical, resulting in the algorithm being flawed and not for use in a practical application.

Contents

1 Introduction

For the simplicity of this report, some standard financial metrics have been assumed or roughly calculated. In a more in-depth analysis, correct data would be collected complete with citations.

A large proportion of money in the economy is created by banks. Banks generate money predominantly through investing, fees and interest on debt. Taking the case of Lisbon's central bank, it has been calculated that the bank can loose as much as €517,000.00 each year due to defaults on personal loans. This is calculated from the assumptions that the 1-Year frequency of default on personal loans is around 3.9% [1] in Portugal and the population of Lisbon is 500,000. It is further assumed for this report, that personal loans are equal to credit.

AI models are only as good as the training data supplied to them. This fact is somewhat over looked or misunderstood by the layman, who is sometimes tasked with data collection. This paper demonstrates the importance of providing unbiased data for model training, while further outlining how to collect such data.

The algorithm devised in this report aims to increase money generated from interest on debt, by reducing rates of default on credit cards. Using the figure given for 1-Year frequency of default on personal loans in Portugal, we can assume that classification models for personal loan defaulters is approximately 96.1% accurate. There is room for improvement regarding the classification of potential defaulters. First, we will discuss the methods we used, namely the dataset [2], pre-processing, analysis and classification models used. This is followed by results of the algorithm, its metrics and the amount of bias discovered in the raw data. Next, an in-depth discussion of the results and the business problems outlined, including data collection methods for reducing bias. Finally, a conclusion summarising the report.

2 Methodology

2.1 Dataset

The dataset was originally used to predict if a customer will make a term deposit from a telemarketing campaigns statistics. There is versatility in the features included in the dataset, that also enable other analysis to be conducted. Categories of dataset features include:

- Personal client data such as age, job type, education, marital status etc.
- Marketing campaign data, which is ultimately removed.
- Contact attributes, again relating to the marketing campaign, which are also mostly removed.
- Social and economical data such as consumer price index, consumer confidence index etc.

The dataset has 21 features including the default labels, with 41,188 observations. The features consist of 9 continuous numerical features and 12 non-numerical categorical or binary features. Initial exploration shown approximately 3000 'unknown' values spread over 5 features. Approximately 25% of the labels were initially labelled 'unknown'. For the use of this analysis, 'unknown' was encoded to 'yes'. The rest of the default feature retained their original values, 'no'. This dataset is therefore theoretical, resulting in the algorithm being flawed and not for use in a practical application.

2.2 Pre-processing

To start, observations that had 'unknown' values were dropped. Non-numerical categorical and binary data was encoded where appropriate. The job types feature was encoded using approximate salaries as job type was deemed a significant feature for this analysis. Certain job types that had no specific salary such as 'retired' and 'self-employed' were encoded with the average salary for Lisbon. This resulted in a salary distribution more closely resembling a right skew which is expected in a population by salary distributions. Using a correlation matrix, features with correlation close to 0 were dropped before continuing.

Next, outliers were removed using z-score normalisation. Each observation that was above or below 2 standard deviations (SD) from the mean were classed as outliers and removed from the dataset. A range of 3 SD was considered, but resulted in many outliers remaining. Furthermore, accuracy of the model increased at a range of 2 SD.

Sampling bias was reduced. This was accomplished by equating the number of default and non-default labels. This ensures that the model is trained over a dataset without generalising to the larger population, thus reducing a potential skew in the results. The model will now correctly identify who will or will not default on their credit. A pre-processed dataset consisting of 10 features and 11,790 observations remained ready for training.

Finally, the dataset was split into training and test sets with an 80/20 split respectively. Standard scaler is used to normalise the datasets. Principle component analysis (PCA) is used to reduce dimensionality of continuous features. The result is two datasets, ready to train the model and test the models capabilities.

2.3 Exploratory Analysis

Correlation matrices were used to clearly identify any correlations between features, mainly, the correlations between features and labels. Count plots were used to visualise categories with high or low proportions of defaulters. Count plots are histograms for categorical data and are very simple to interpret. Categories visualised include marital status, education and job. Further count plot visualisations were used for the age feature, with the aim of identifying ages more or less likely to default.

Outliers were analysed to ensure they were not significant to the analysis. This was accomplished by calculating the ratio of defaulters in the outlier dataset and original, unprocessed dataset. The ratios were then compared. Nearly a third of defaulter labels were removed which is highly undesirable. This had no effect on the accuracy of the model due to size of remaining dataset. Equal ratios indicated that the outliers and original dataset had equal chances of default.

2.4 Extreme Gradient Boost Classifier

The XGB classifier implements gradient boosted decision tree design to reduce run times and improve performance. Due to the mixture of numeric categorical and continuous data, such a large number of observations for training the model and its relatively low run time, the XGB classifier performs well in this classification task. A RandomForest classifier was compared, but due to such large run times was discounted. The accuracy of classification is the most important output of the model, but the XGB classifier also boasts a relatively low run time. The XGB classifier therefore produces the best accuracy to run time ratio for this problem.

3 Results

3.1 Exploratory Analysis

Figure 1 shows the correlation matrix of the processed dataset. The social and economical features have the greatest magnitude of correlation, followed by the client data feautres. The job feature showed a lack of correlation to the label which was surprising. However, this was anticipated due to the need to encode approximate salaries.

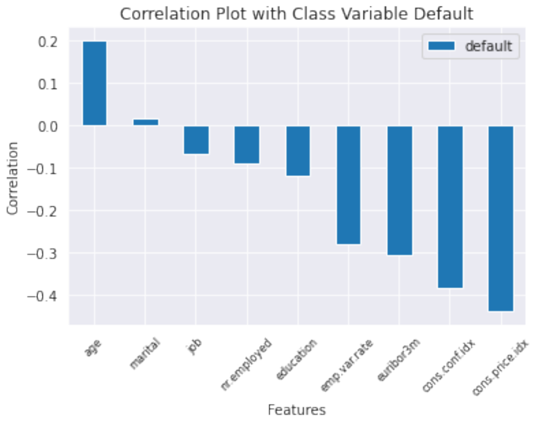


Figure 1: Graph to show correlation of continuous variables with default label.

Figure 2 shows that those with lower salaries are more likely to default. It is further shown in figure 2 that higher educational levels result in lower rates of default. There is no clear relationship between marital status and default rates, as expected.

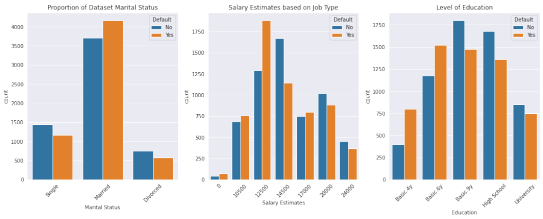


Figure 2: Ratio of default for categorical variables.

Figure 3 shows an age distributions. For the customers in the age range 17-34, around 38% have defaulted. The age range 34-40 shows an almost equal likely hood of default. Beyond the age of 40, nearly 60% of customers had defaulted.

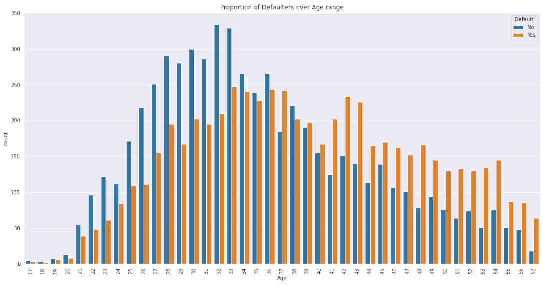
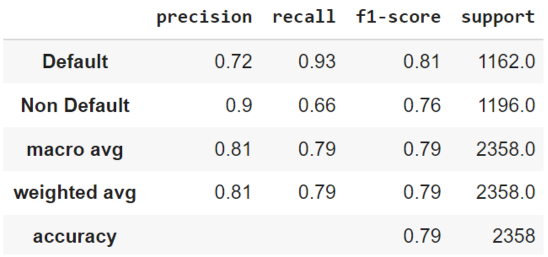


Figure 3: Ratio of default for ages ranges from 17-57 (blue bars not default, yellow bars default).

The original dataset and processed dataset both had ratio of defaulters close to 20%. The ratio of defaulters in the outliers was 22%. This shows the outliers removed were reflective of the original default ratios, therefore having minimal effects on the analysis.

3.2 XGB Classifier Metrics

Figure 4 shows a summary of the classification metrics achieved using the XGB model. Regarding defaulters, the model has an accuracy of 0.79, a precision, recall and f1-score of 0.72, 0.96 and 0.82 respectively. The support for the test is 2358 in total. Further metrics are shown, including non-default classification metrics.



	precision	recall	f1-score	support
Default	0.72	0.93	0.81	1162.0
Non Default	0.9	0.66	0.76	1196.0
macro avg	0.81	0.79	0.79	2358.0
weighted avg	0.81	0.79	0.79	2358.0
accuracy			0.79	2358

Figure 4: Classification report for XGB model

To easily visualise the misclassification decisions the algorithm has made, a confusion matrix is given by figure 5. It can be seen that a large proportion of default prediction are correct, non-default predictions are lower.

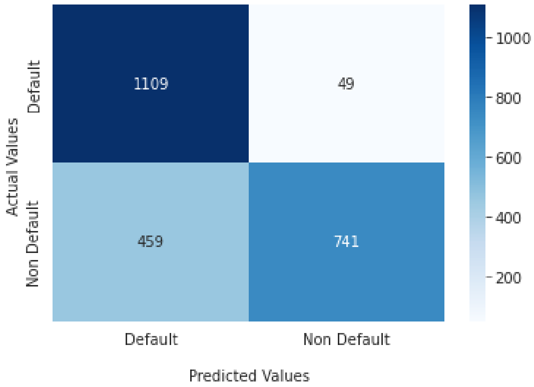


Figure 5: Confusion matrix of classification model. Greater magnitudes are highlighted by darker blue colour.

3.3 Algorithmic Bias

Sampling bias has been reduced as much as possible for this dataset. There still exists a large amount of bias however, greatly effecting the accuracy of this algorithm. The magnitude of which is unknown and incalculable due to flaws in the raw data provided. This form of bias cannot be mitigated. The outcome is a large skew in the results of the classifications made by the algorithm, reducing accuracy.

4 Discussion

4.1 Analysis

There are some expected patterns being shown by the count plots. For example, it is expected that those with lower salary are more likely to default than not, this is shown. Furthermore, those with higher educational levels default less, this is likely due to them having greater salaries.

The differences in default ratios per age range could be due to a number of factors. Firstly, those in the age range 17-34 could be less likely to be accepted for a credit card, or even credit they could not afford to pay back. Then, as age increases, the likely hood of living through a financial market crash or even financial instability could increase. Furthermore, as age increases there is responsibility placed on a customer in terms of dependencies (children, elderly relatives).

4.2 Credit Card Default Prediction Algorithm

The algorithm identifies potential defaulters with 79% accuracy. An accuracy of 79% is below what is required for an improvement on the current model. Algorithmic bias likely has the largest negative effect on the algorithms accuracy.

The algorithm has a precision value 0.72. The model therefore identifies 72% of customers as potential defaulters. The bank might lose out on potential credit card customers here due to a relatively low precision value. Precision is therefore the indicator of missed potential profit.

The model correctly identifies 96% of defaulters identified. Of the potential credit card customers identified, the algorithm may allow 4% of potential defaulters to be granted a credit. This is undesirable for a bank as it likely amounts to a very large loss. The recall is the most important statistic for this financial problem, as giving credit to a defaulter is probably more costly than not awarding credit to a customer who is unlikely to default. Recall is therefore the indicator of potential loss to default.

4.3 Algorithmic Bias Reduction

Algorithmic bias likely has the largest effect on the accuracy of the algorithm. There exists algorithmic bias in the model due to the lack of data regarding people who were not accepted for a credit card. The model is only trained on people who qualify or are given a credit card creating a large bias and skew in the results. To create a model that can identify defaulters with a higher accuracy, a dataset with information regarding people refused credit must be supplied. Therefore, it is recommended that when gathering data, to also gather data of customers who have been refused credit. Data gathered must be representative of the real world for this particular case.

5 Conclusion

The chosen classification model yielded an accuracy of 79% with a recall of 96%. This is below the accuracies of current models of default classification. Never the less, this project gives great insight into the problem of bias effecting machine learning algorithms, with real world outcomes. Such an example could be a defaulter potentially being given a credit card. This report shows a great example of how to collect data so that it is representative of the environment the algorithm is designed to work in and is further reflective of the problem at hand.

Due to the algorithmic bias present in the data used to train this model, it is not suitable for practical applications. However, such a high recall results in most potential defaulters being identified. The 4% of defaulters not identified may cause too much financial loss for this algorithm to be effective. If the raw data was more representative of the problem, this algorithm could likely achieve desirable accuracies and metrics.

The XGB classifier used has excellent accuracy with a relatively low run time. For this reason, it is a highly desirable classifier for quickly and accurately deciding if credit should be given. It is expected that with a large training dataset consisting of 9432 observations, run times would be longer.

References

[1] Jean Dermine and C Neto De Carvalho. Bank loan-loss provisioning, central bank rules vs. estimation: The case of portugal. *Journal of Financial Stability*, 4(1):1–22, 2008.

[2] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.