Evaluating the Performance of Classification Algorithms in Predicting Credit Card Default

CIND 820: Big Data Analytics Project

Literature Review, Data Description and Project Approach

Mitchell Cassar

501140751

Dr. Ceni Babaoglu

February 20, 2023

This report seeks to expand on the proposal of the *Abstract* of this project. This is accomplished in two ways. Firstly, a comprehensive review of the existing literature of this topic is provided. Background context of this study is explored and summaries and conclusions of previous work from an array of sources are analyzed and evaluated. Secondly, an in-depth study on the data used in this project is conducted including a data dictionary, summary statistics and initial exploratory analysis. These two tasks combine to set the stage of this project, attempting to examine how the selected data can be used to perform a useful analysis that fits in the existing landscape.

Literature Review

Considering the immense financial and regulatory incentives pertaining to accurately predicting customer behaviour, it is not surprising that there exists a great deal of literature on this topic. Academia, private sector firms and even sovereign governments have all contributed reliable research attempting to answer similar questions as this report. In fact, many nations have gone as far as to *legislate* the mandatory publication of materials on this topic for the purposes of public interest. As an example, the most recently enacted version of *International Financial Reporting Standards* requires financial institutions to include a provision for Expected Credit Losses in shareholder reporting (OSFI, 2016).

Assigning a value to an entity's creditworthiness is a well-known phenomenon. While the vast financial reporting requirements mentioned earlier are of concern to businesses, this topic has a ubiquitous impact on the lives of the public as well. The process undertaken in this project (attempting to design a system which can accurately forecast a borrower's future default) appears

in everyday life in the concept of a credit score. Credit scoring is common all over the world, where a person is assigned a creditworthiness value based on different demographic and financial factors. The information age has only enhanced the importance of this process. Although credit scoring is controversial, it performs well. A 2022 study of over 20,000 Chinese personal loans classified personal loan defaults with and without credit score included (Wang et al., 2022). It concluded that the inclusion of credit score positively impacted classification accuracy for all 5 algorithms tested (Wang et al., 2022).

While credit scoring attempts to deduce creditworthiness, this project focuses specifically on credit card defaults. Several studies have attempted to answer this exact question in the past. Studies on this topic range in geography, datasets, variable inclusion, algorithm selection and evaluation metrics. However, the conclusion is nearly universal: machine learning algorithms are an effective predictor of binary future default. In 2022, Achsan et al. found that for over 100,000 Indonesian records, several demographic characteristics such as age, sex and geographical location played a statistically significant factor in predicting default using a logistic regression algorithm (Achsan et al., 2022). This study is notable for combining both demographic and behavioural variables to reach its conclusion, but only uses a snapshot of payment history rather than a time-series (Achsan et al., 2022).

Yi et al., in a 2019 Chinese study used a proportional hazards model analysis method to notably conclude that borrower income was not related to credit card default (Yi et al., 2019). Among their other findings are that credit card customers who engage in more online purchases are more likely to avoid default (Yi et al., 2019). This study also suggests high importance of geographical and demographic variables (Yi et al., 2019).

Another notable study in this area is that of Chen and Zhang in 2021. This study tested a dataset with 6 different machine learning algorithms but focused on an unusual metric: model performance (Chen & Zhang, 2021). This study concluded that the support vector and logistic regression algorithms performed significantly slower while maintaining a similar accuracy to other methods (Chen & Zhang, 2021). While model performance is not the most important factor in approaching this topic from an academic perspective, for an institutions production environment which must quickly analyze potentially millions of records it is a critical variable.

Studies have even been conducted using this exact dataset. A 2019 study by Teng et al. used a process very similar to this project: comparing the performance of various machine learning models (Teng et al., 2019). While their research supports the findings of other studies of the landscape, this study is important to focus on due to the evaluation metrics used (Teng et al., 2019). This study used only accuracy to compare the performance of the algorithms (Teng et al., 2019). Accuracy is the proportion of records whose class is correctly identified (Handelman et al., 2019). Although a useful metric, accuracy alone does not validate the performance of a model. It does not account for, for example, false positives. In this case, these borrowers would be labelled likely to default when they are not. This project will explore the real-world implications of these situations using various evaluation metrics.

By combining the literature cited with industry research, the overall context of this project can be set. When setting the stage, the first question worthy of consideration is the following: what is known about this area? Based on the findings of the works previously cited, there are a few conclusions that can be drawn to form a consolidated view of the landscape. Firstly, credit scoring is confirmed to be a statistically significant way to reduce financial risk for lenders. This is confirmed not only by the research, but also intuitively by observing the time and

resources that go into improving this process for financial institutions. Next, it is confirmed by the existing literature that machine learning algorithms can improve the accuracy of this process. It is important to note that there is no consensus conclusion on a more granular level (e.g. which algorithm performs the best), but essentially “any algorithm” performs better than “no algorithm”. This conclusion holds for the differences that vary between studies, including location, variables used and evaluation metric. Essentially, the research suggests that as long as basic personal information and credit card payment history are used, a machine learning model will be able to improve default prediction capabilities.

Before expanding on the justification and approach of this project, it is important to critically analyze the existing literature. The most basic critique of these studies is ethical. The impacts of credit scoring are felt throughout an individuals entire life, and the outputs of these models determine if and how they are able to do things like purchase a house. In the example previously cited, a false positive classified by a model may end up being someone denied credit for which they are able and willing to repay. The variables used in these analyses are important to consider as well. To what level should demographic data be used to improve classification? If, for example, gender is analyzed to be a significant predictor of credit default, should it be used to forecast future applications? The Canadian Human Rights Code forbids discrimination of the basis of areas such as sex (Government of Canada, 2023). Is an institution potentially liable for the recommendations of machine learning models?

The second critique of the existing literature involves the inconsistency of the conclusions drawn. Although they all come to the same broad conclusion, more narrow conclusions differ greatly. Among the many similar studies already conducted, there is huge variance in which variables matter, the correlation of attributes, the weighting of demographic

vs. financial data etc. The results of these studies are extremely dependent on the specific dataset studied. Studying credit repayment is studying human behaviour, in which all factors and variables cannot possibly be accounted for. In a landmark 2005 study, John Ioannidis suggested that most published studies suffer from a "crisis of replication" (i.e the results do not reappear in subsequent studies) (Ioannidis, 2005). Thus, it becomes evident that the results of this study should be evaluated in this context, and should not extrapolate its findings too broadly without first proving correlation.

Now that the existing literature has been noted, analyzed, and critiqued, it is possible to evaluate where this project resides in the current landscape. Considering this topic is well studied with consistent conclusions, it is fair to wonder what ground this study is attempting to break. This study will closely resemble the work of Teng et al. previously mentioned. The similarities and differences between that study and this proposed project are noted below.

Like Teng's work, this project will use multiple machine learning algorithms to analyze an open-source dataset of credit card customer data to compare their results. This process includes, in chronological order: initial data analysis, feature selection, algorithm application and model evaluation. The dataset in common is a collection of 30,000 records of Taiwanese credit card customers and includes both demographic and prior financial behaviour information. This dataset is available from the University of California-Irvine machine learning repository and is curated specifically for evaluating models (Yeh, I.C., 2019). More information on the dataset can be found in the *Data Description* section of this document.

Another similarity between these works is the algorithms selected. The work cited uses the following methods to classify records:

- K-Nearest Neighbours
- Decision Trees
- Boosting
- Support vector machines
- Neural Network

While this project will also make use of the K-Nearest Neighbours and Decision Tree algorithms, the remaining methods will be considered out of scope. Instead, this project will introduce Logistic Regression and Random Forest algorithms to the equation and evaluate their performance alongside the previously studied methods.

This project will also differ in the second step of the data analysis roadmap: feature selection. Teng's method for selecting the optimal set of features is results-based. They simply selected the optimal combination of parameters that produced the highest accuracy in model production (Teng et al., 2019). This study, however, will perform feature selection *before* training the models, using traditional mathematical methods such as removing highly correlated or low variance attributes.

The last main difference of note between these two works is the selected evaluation metrics. Teng's study evaluates various methods on accuracy only (Teng et al., 2019). As previously mentioned, accuracy measures only the proportion of records assigned the correct class. This study will use a more comprehensive evaluation technique, measuring precision, recall and F-score in addition to accuracy. A confusion matrix will be provided for each algorithm, and the real-world implications for each of these measures will be explored. For the purposes of this report, the 'best performing' algorithm will be considered that which balances successful record classification and error minimization rather than the former alone.

To summarize, although this topic and this specific dataset have both been previously studied, there is still room for exploration. In essence, this dataset has yet to be studied on this collection of algorithms and evaluated using this combination of metrics. This report aims to provide a more comprehensive view of this topic than the existing literature.

Data Description

This section will provide an initial, exploratory analysis of the selected dataset. All actions referenced are undertaken using the Python programming language. Please use the following link to access a repository of the codes used to create this work: https://github.com/mitchell-cassar-torontomu/cind820. Please consult the 'data_description.ipynb' file for the codes used in this step. The dataset used can be found here: http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.

Before any analysis can be conducted, the dataset and its contents need to be stated. As previously mentioned, this dataset contains a combination of demographic and payment history information for 30,000 Taiwanese credit card customers. A data dictionary of the attributes, their data type and description can be found in Appendix I (Yeh, I.C., 2019).

The first transformation performed on the data is to amend the column names. The multi-level index is replaced with clean column names. Next, the data types are cleaned as well. In this dataset, the factor variables are all given as integers, which are interpreted by Python/pandas as such. The relevant factor variables such as age, gender, education and all of the 'pay' attributed are converted to their proper categorical forms. The target variable, "default_payment_by_month", is also amended to be a categorical variable. For the next piece of analysis, the distribution of the binary classification variable is calculated and visualized.

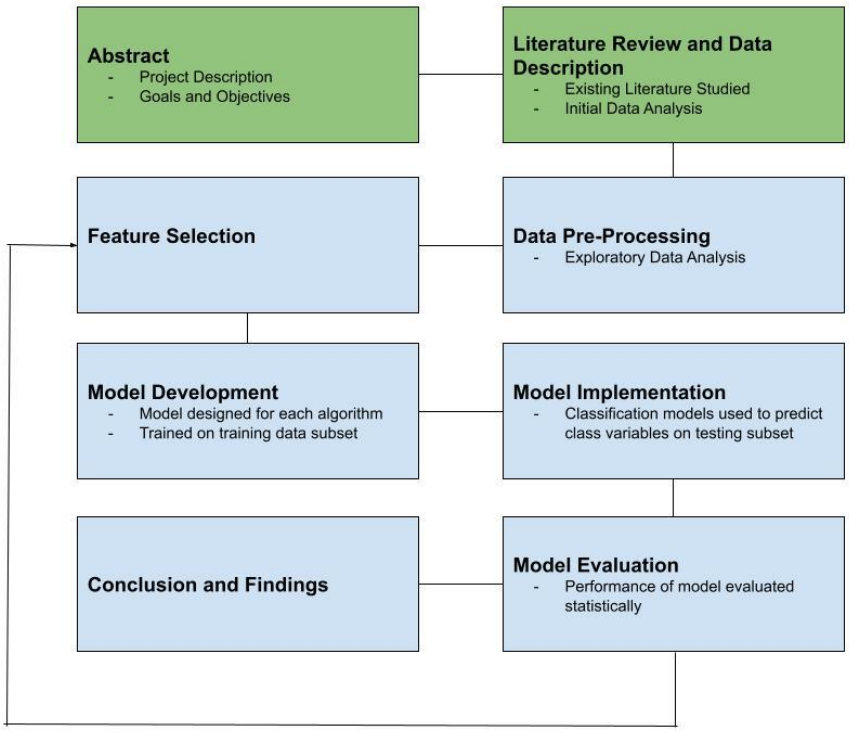
Approximately 78% of the records in the dataset are non-default records (0), while 22% are default records (1). The distribution of the target will be further explored later in this project. Additionally, the dataset is scanned for NULL values. It does not contain any such values.

At the column level, some basic analysis is also offered at this point. The descriptive statistics of each column are provided, including count, mean, standard deviation, minimum and maximum. Again, the results of this analysis will be evaluated in subsequent deliverables.

Lastly, the correlation of attributes is calculated. This step is done using a correlation visualization, to visually identify attributes that are highly correlated. These attributes are obvious candidates to be dropped during the feature selection revisions of the models created in this project.

Project Methodology

The following chart represents the workflow and overall methodology of this project:



This visualization provides a clean overview of the approach to this study. Steps in green-coloured boxes have been completed to this point. The remaining steps follow the standard practice for machine learning analysis: the data is explored, models are developed and trained, and then tested and evaluated on their performance. It is important to note that model development, implementation and evaluation combine to form an iterative process. This is symbolized by the looping arrow from model evaluation to feature selection. These steps will be repeated, where models will be “fine-tuned” with different combinations of parameters. Conclusions and findings of this report will take into account all of the iterations performed and will attempt to contextualize them appropriately.

Appendix I: Data Dictionary of *default of credit card clients* dataset

Source: Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Attribute	*Data Type*	*Description*
Amount of Given Credit	Integer	
Gender	Factor	1:Male, 2:Female
Education	Factor	1: Graduate School 2:University 3: High School 4:Others
Marital Status	Factor	1: Married 2: Single 3: Others
Age	Integer	Age in years
History of Past Payment	Factor	Attributes 6 – 11 each represent a customers payment history for a given month The measurement scale for the repayment status is: -1: pay duly; 1: payment delay for one month; 2: payment delay for two months; . . .; 9 = payment delay for nine months and above. 6: Repayment status in September 2005 7: Repayment status in August 2005 Etc.
Amount of Bill Statement	Integer	Attributes 12 – 17 each represent a customers bill statement amount for a given month 12: Customer bill amount in September 2005 13: Customer bill amount in August 2005 Etc.
Amount of Previous Payments	Integer	Attributes 18 - 23 each represent a customers payment amount for a given month 18: Customer bill payment in September 2005 19: Customer bill payment in August 2005 Etc.
Default Next Month	Factor	The target variable Binary classification variable 1: Customer default next month 2: Non-default for customer next month

References

Achsan, W., Achsani, N. A., & Bandono, B. (2022). The demographic and behavior determinant of credit card default in indonesia. *Faculty of Economic and Business, UIN Syarif Hidayatullah Jakarta*. doi:10.15408/sjie.v11i1.20215

Canadian Human Rights Act. Government of Canada. (2023, February). Retrieved February 12, 2023 from https://laws-lois.justice.gc.ca/eng/acts/h-6/page-1.html

Chen, Y., & Zhang, R. (2021). Research on Credit Card Default Prediction Based on k-Means SMOTE and BP Neural Network. *Complexity*. https://doi.org/10.1155/2021/6618841

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Guy S. Handelman, Hong Kuan Kok, Ronil V. Chandra, Amir H. Razavi, Shiwei Huang, Mark Brooks, Michael J. Lee, and Hamed Asadi American Journal of Roentgenology 2019 212:1, 38-43 https://www.ajronline.org/doi/full/10.2214/AJR.18.20224

IFRS 9 Financial Instruments and Disclosures. Office for the Superintendent of Financial Institutions. (2016, June). Retrieved January 28, 2023 from https://www.osfi-bsif.gc.ca/Eng/fi-if/rg-ro/gdn-ort/gl-ld/Pages/ifrs9-22.aspx

Ioannidis J. P. (2005). *Why most published research findings are false.* PLoS medicine, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Li, Y., Li, Y., Li, Y. (15 July 2019). *What factors are influencing credit card customer's default behavior in China? A study based on survival analysis*. Physica A: Statistical Mechanics and its Applications, 526. https://doi.org/10.1016/j.physa.2019.04.097.

Teng, H., Lee, M. (2019). *Estimation Procedures of Using Five Alternative Machine Learning Methods for Predicting Credit Card Default.* Review of Pacific Basin Financial Markets and Policies, 22(3). https://doi.org/10.1142/S0219091519500218

Wang, H., Chen, W., & Da, F. (2022). Zhima credit score in default prediction for personal loans. *Elsevier*. doi:10.1016/j.procs.2022.01.188

Yeh, I. C., & Lien, C. H. (2009). *The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients*. Expert Systems with Applications, 36(2), 2473-2480.