EasyChair Preprint

№ 1760

A Hybrid Machine Learning Model with Cost-Function Based Outlier Removal and Its Application on Credit Rating

Aurora Mu

October 24, 2019

A Hybrid Machine Learning Model with Cost-Function Based Outlier Removal and Its Application on Credit Rating

Authors: Aurora Mu

School: Western Connecticut State University

Danbury, Connecticut, United States of America

Advisor: Xiaodi Wang

August 2019

ABSTRACT

With the rapid growth in the credit industry, the ability to allocate capital efficiently and profitably is of great significance to financial institutions. Banks and credit companies often have sizeable loan portfolios, making it necessary to develop accurate credit scoring models. Slight improvement in credit scoring accuracy can reduce lenders' risk and translate to significant future savings. Machine learning techniques such as support vector machines, neural networks, and logistic regression learning, are widely explored and utilized. In this paper, using Lending Club loaner information data as dataset and credit rating as subject, we explore a hybrid machine learning methodology, which combines different algorithms in different stages of data processing, training and prediction. In the data preprocessing stage, we introduced a cost based outlier removal technique which can generalized to all types machine learning algorithms. In our experiment, we implement logistic regression during feature treatment to reduce feature dimensions and the sample cost of a particular machine learning algorithm are calculated as the basis for outlier detection and removal. We create three models of support vector machine (SVM), decision tree (DT), and logistic regression (LR), and three hybrid models incorporating our new ideas into SVM, DT, and LR. The traditional and hybrid models are compared by efficiency, F_1 score, accuracy, recall, AUC, and precision. The results demonstrate performance improvement of the Hybrid models.

Keywords — *Machine Learning, Outlier Removal, Credit Score Modelling, Hybrid Method, Support Vector Machine, Logistic Regression, Decision Tree, Cost Function*

HIGHLIGHTS

We introduce a new **cost-based outlier removal** algorithm as a step of preprocess of training data, we implemented a hybrid machine learning algorithm, aiming to combine the power of different machine learning algorithms on different types of features and hypothesis. We experiment combination of three types of machine learning algorithms SVM, DT and LR. The new hybrid models shows improvement in performance compared to the traditional SVM, DT, and LR. This new methodology can be further explored with other algorithms and applications.

BACKGROUND

With the rapid growth in the credit industry, the ability to allocate capital efficiently and profitably is of great significance to financial institutions. Banks and credit companies often have sizeable loan portfolios, making it necessary to develop accurate credit scoring models. Even the slightest improvement in credit scoring accuracy can result into significant future savings, reducing lenders' risk [6].

The main objective in credit scoring models is to measure the creditworthiness of borrowers. In other words, the model determines whether the borrower is at risk of defaulting based on predictor variables such as income, FICO score, and loan interest rate. Creditors typically construct models based on preexisting data with relating predictors to default rates. Thus, when presented with a credit or loan application, banks can assess the creditworthiness and accept or reject the applicant. For creditors, the credit score models are now not only used for credit loan decisions, but also for risk management. Therefore, the push for better credit score models has led to development and research into various techniques, particularly in machine learning, as credit scoring is well suited for supervised learning [7].

The credit scoring problem can be approached through a machine learning framework structured in the following way. Given a set of debtors, each debtor has a number of characteristics (i.e. annual income, debt-to-income ratio, FICO score, etc.) called predictor variables. Given a dataset $S = \{(x_1, y1), (x_m, y_m)\}$ where m is the number of samples (borrowers), each borrower x_i has n attributes. y_i indicates status (defaulted or not). This dataset is used to create model f. Then for an unknown x, which would be a loan applicant, we assign the default status y as y=f(x). The output y is binary, with 1 denoting default (charged off) and 0 denoting normal (paid in full). This framework, in which a model identifies a binary class based on new observations is known as a classification problem in machine learning literature [9].

Many different binary classification techniques have been explored for credit scoring. In particular, the most common methods are: neural networks (NN), support vector machines (SVM), fuzzy logic, logistic regression (LR), discriminant analysis, decision trees (DT), and Bayesian networks [9]. In addition, hybrid methods can be used as well, which involves combining different tradition techniques to improve the performance and accuracy of the model [9]. For example, a combination of discriminant analysis and backpropagation neural networks can be used to evaluate credit scoring [8].

In credit score modelling, there are a variety of machine learning techniques. Neural networks have seen the most research [9], closely followed by hybrid methods. West [15] compared five neural network models with traditional models, concluding that credit scoring accuracy is indeed improved through neural networks. In addition, it suggested that logistic regression is a good alternative. However, the debate over which technique is best for credit models is widely argued, with conflicting data arising. For example, Desai et al [3] concluded that neural networks performed significantly better than linear discriminant analysis, whereas Yobas *et al* [16]

claimed the opposite. Furthermore, Louzada [9] noted the absence in research for discriminant analysis after 2006, and fuzzy logic after 2010.

Other proposed methods involve ideas into feature selection and creation of models. Since credit data often contain many features that would be irrelevant or redundant, feature selection is a common practice in credit modelling. Waad et al [14] propose a three-stage feature selection utilizing a variety of filter and wrapper methods to reduce irrelevant features. The conclusion showed that the new method of feature selection produced either superior or adequate results compared to benchmark methods. Somol *et al* [11] compares filter and wrapper methods. Chen et al [2] propose a SVM-based method for feature selection. Chai and Chen [13] look into combining clustering and classification techniques and compares them, concluding that a combination of classification techniques performs the best.

It is noted that the accuracy of the models tend to stay in the 80% percentile range, rarely going over 90% [9]. Hence the continued research and development into credit scoring models, as the need for accurate models is ever present.

RESEARCH OBJECTIVE

The goal of this research is to establish a hybrid machine learning model and as one application, test it on credit score issue. SVM, DT, and LR are used as benchmarks in comparing our model. We implemented a logistic-regression-based learning method during the data preprocessing to treat categorical variables and similar features. Since the values in the categorical features are normally discretely assigned by human with little consideration into how these values are related to the target variable, those values might not truly reflective of their weight in the model. Similar features report similar but slightly different information; however, the removal of one of them may result in loss of valuable information. After processing both the similar and categorical features through LR, a score is obtained for each sample in the treated features that reflects its importance for either categorical feature or combination of similar features in the model. We also introduce a cost-based outlier removal method in the data pre-process to improve the performance of the model. We then compared the hybrid models to the no- hybrid models.

DATASET

We used loan data collected from Lending Club's Loan Origination Data [17], a US peer-to-peer lending company that is currently the world's leading peer-to-peer platform [10]. Lending Club enables borrowers to apply for loans. Loaners can choose to accept or reject loans based on the information provided on the listing. The data covers all loan information from 2007 to 2018, and has 150 variables and over 2.26 million samples[1], as well as the final loan response (loan status paid off or default). However, there are features which may contain missing values, features are missing over 30% will be removed. In addition, only feature which would be known to investors

[1] Exact number is: 2,260,701

are limited to the applicant's information available on the Lending Club loan listing, so the variables available are taken into consideration. More on feature selection and data processing will be discussed in the next sections.

METHODOLOGY

A traditional machine learning models are generally broke down into following steps: Data collection, feature selection, data preprocessing, train model and validation, test model, and evaluate accuracy. There are various techniques in each steps. In our model, the architecture is shown as follows (Figure 1)

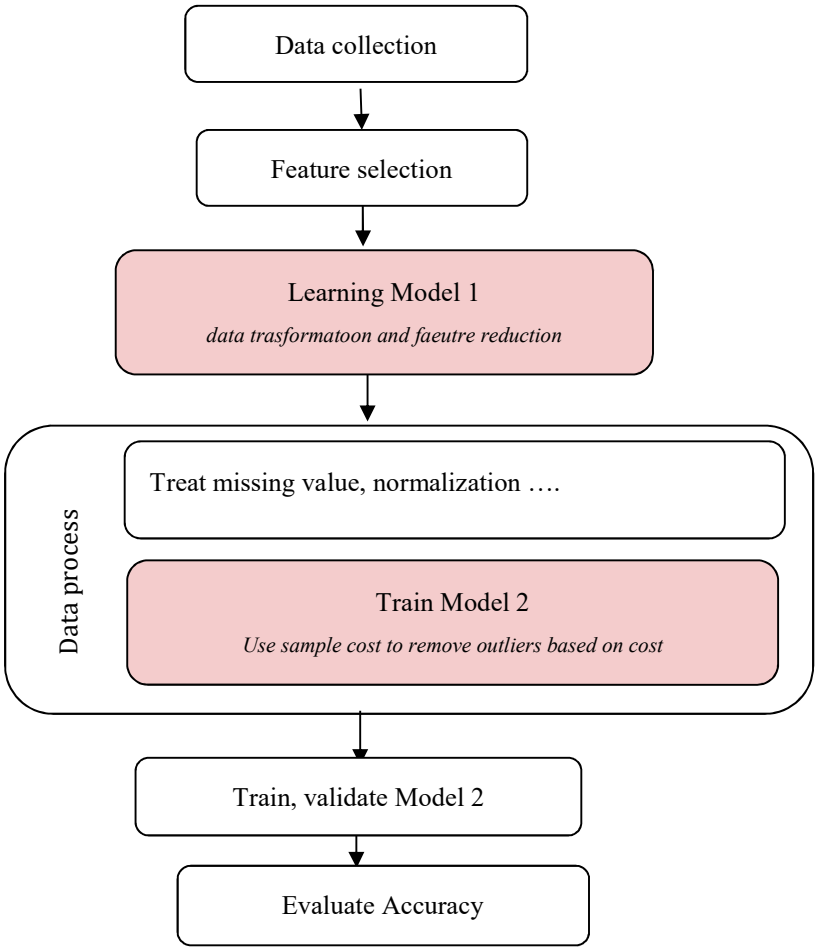


*Figure 1: O*verall logic structure, where red boxes indicate our additions

As shown above, we create a hybrid model that, after initial feature selection, utilizes a Learning Model 1 (could be more than one model) to transform data of categorical features and combine similar features. The original dataset, defines as Dataset 1, is converted into a new dataset,

defined as Dataset 2, with some new features whose values are either the logistic regression cost results of one categorical feature, or the logistic regression cost results of a group of similar features. Dataset 2 has no more discrete categorical values and a reduced number of features. Then for comparison purpose, the samples of Dataset 1 and Dataset 2 are split into training, validation and test set in same way. Dataset 1 is processed in the traditional manner for machine learning: treating character and missing values, and normalizing data etc. For Dataset 2, however, will go through an additional cost-based outlier removal process which is described as follows.

Once we choose the machine learning model (Model 2) for final prediction, no matter it is SVM, LR, DT, or other models, we use the training data of Dataset 2 to train this model for the first round, then calculate the cost of each sample in the training set, use standard deviation of the sample cost as a criteria to detect outliers. After the outliers are removes, Dataset 2 becomes a Dataset 2' which has less number of samples in its training set. Then

In this study, we use linear regression as learning model 1, and test SVM, LR and DT as learning model 2 individually. The following sections will discuss the steps of our methodology in detail. The programs are implemented with Python

DATA PROCESS

Sample selection

The raw data comprises information on borrowers and their loan statuses: "current", "fully paid", "charged off", "late", etc. Since the focus of this model is to determine between "good" loan applicants ("fully paid") and "bad" or risky applicants who defaulted on their loan ("charged off"), we filtered for those two statuses, creating a data set of 1.3 million[2] loan samples.

Feature selection

The first step in feature selection is to remove unnecessary features, such as applicant's website url ('url'), or membership ID number ('member_id'). These features don't have an effect on the credit score, as they serve as basic information for the lenders to view.

At Lending Club, borrowers can receive loans by putting a loan listing. Investors than can search and select loans based on the information provided. The original data contains some features which are not accessible to the investor during the loan application. Those features are excluded from our dataset.

The next step treat missing values. Features which miss over 30% of their data are excluded. The other missing values are replaced with average value of the feature.

In next step, we removed redundant features that overlapped with other categories. For example, the information for loan grade is incorporated in the loan subgrade, creating repetition.

---

[2] Exact number is: 1.345,710

The feature 'grade' has categories from A-G (i.e. C), however 'sub_grade' is a further breakdown of grade, assigning a number from 1-5 to each letter (i.e. C2). Therefore, 'sub_grade' already contains the information from 'grade' and in fact is much more specific. Several other features too are repetitive in this manner, so we removed them.

The raw loan data set from Lending Club initially had 150 variables. After feature selection, 36 variables remained. The reduced data is denoted as Dataset 1.

Assign values to categorical features

The 36 variables in Dataset 1 are denoted with a mix of character and numerical values. We assign a value to each group in a particular categorical feature. For example, 'loan_status', which is our response column, is formatted as "Fully Paid" or "Charged Off"[3]. We assigned {1} for charged off, and {0} for fully paid.

Logistic-regression based feature treatment

Hybrid methods combine different machine learning techniques in order to improve the performance of the model. In our model, we use logistic-regression-learning algorithm as learning model 1 and combine it with other models. When reexamining the remaining 36 features, several features contain similar information. For example, one feature lists the number of revolving accounts while another lists the number of revolving accounts with a balance greater than zero. Although these variables could be redundant, removing one might result in loss of valuable information. However, keeping both might lead to weaker performance of the model, we combine features with similar information by training an LR model. The LR model returns the probability of being classified as 1 for each sample. These probabilities are our combined "score" for the two predictor variables. The new column of the probability score is used to replace the original two features. In addition, when converting categorical variables into numerical variables. For example, the states of the applicant are assigned a number from 1-51, depending on the alphabetical ordering. Alabama is assigned as 1, Alaska as 2, Arizona as 3, etc. However, these numbers have no intrinsic meaning and do not reflect a state's influence in the model. Therefore, in a similar manner we extract the target variable and used LR learning to obtain a probability scores, which replace the original numbers and have the probability information encoded in them. The resulting dataset is denoted as Dataset 2.

Treating missing values

The absence of values in dataset is a recurrent issue that may be a result of several issues such as human input error or recording mechanism malfunctioning. Prior to applying a credit scoring method, it is critical to pre-treat the data, as missing values will throw off the learning

model. One possible approach is to simply ignore and drop the missing numbers from the original dataset. Another way is to replace NaN values, whether through mean substitution, or regression. In this paper, mean substitution is utilized; the mean of a particular feature is found, the NaN values of that particular feature are replaced with the calculated mean.

Normalizing data set

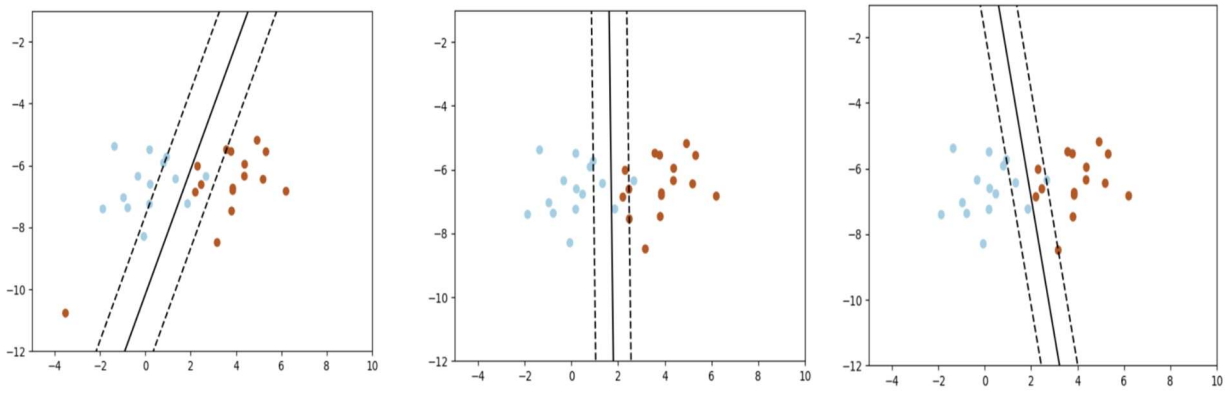
Financial organizations collect a wide variety of consumer variables, which can have drastically different ranges in the raw data. For example, age can range from 0-100, whereas yearly income can range from 0-100,000, or higher. Due to the larger raw data values, the machine learning model might improperly view yearly income as more significant. Thus, after treating error values in the dataset, it is necessary to normalize the data. Several methods of data normalization can be used, such as standard score (z-score), and min-max. In this paper, min-max normalization is used.

Outlier removal and class weight

After the data normalization, we introduce a cost-based outlier removal process, which evaluates the cost function of each sample, and removes samples based on a determined threshold. The cost function measures the error in the model's prediction, expressed as the difference between the predicted and actual value [9]. Machine learning models aim to minimize the cost function in order to improve performance. For LR and DT, we set a threshold at 95% such that a sample is labeled and removed if it is wrongly classified with a possibility greater than 95%. For SVM, we set the threshold at 2.3σ ($\sigma$is the standard deviation of the cost of all the samples) such that if a wrongly classified sample is removed if it's cost is over 2.3 σ. Once the samples are removed, we reset the models and retrained using the reduced training set.

Lending Club's dataset is a skewed dataset with about 80% of the data had the class label $\{0\}$, the imbalanced data can cause learning models to label too many samples as $\{0\}$. Therefore, by adding class weight to the data, the model will penalize the misclassifications of the minority class more heavily, resulting in an increased true positive rate [12]. In this study we adjust class-weight to find an optimum weight about 1.8 on the minority class.

To demonstrate the effect for this outlier removal process and class-weight, we construct a two-variable classification problem and Figure 2 illustrates process with SVM as learning model.



(a) SVM with an initial outlier (b) SVM after removing outlier (c) SVM adding class weight

Figure 2

Figure 2. (a) shows a SVM model with an initial outlier of red dot in the lower left corner. The hyperplane is evidently skewed towards the outlier, causing several blue inputs to be misclassified, as the hyperplane attempts to incorporate the red outlier. After executing the cost based outlier removal procedure and train the SVM again, In Figure 2. (b), we can see how the hyperplane becomes more vertical and separate the two groups better. While the model is improved, there are still misclassified points. Figure 2. (c) shows that after adding class weights, there is one less misclassified blue input.

TRAINING AND TESTING

During the training process, 20,000 samples are randomly selected from both Dataset 1 and 2. The training set is then split into 80% training and validation, and 20% testing. We ran 50 trials, where each trial created three traditional modal of LR, SVM and DT and three new models with LR, SVM and DT as predicting model but have our modification integrated in. After 5-fold cross validation. The models are run on test set

RESULTS

After each trial, we collect the test metrics of each hybrid model and compare them with those from unmodified model. A typical result of one trial is shown ad follows, where the postfix "–H" denotes the result from hybrid model, "-O" denotes the results from unmodified model

Run Index	Model	Training time	tp	tn	fp	fn	# of Outliers
3	DT-O	0.3675	517	2938	265	280	0
3	DT-H	0.295286	527	2970	256	247	176

3	LR-O	0.202038	458	3076	131	335	0
3	LR-H	0.169056	604	2975	214	207	40
3	SVM-O	7.602738	559	3018	189	234	0
3	SVM-H	7.099988	651	2914	279	156	234

Table 1, Training time, TP, TN, FP, FN and number of outliers of one trial

We collect training time, TP, TN, FP, FN and number of outliers removed by hybrid model, then calculate each model's accuracy, precision, recall, F1-score and AUC values

Run Index	Model	accuracy	precision	recall	F1-Score	AUC
3	DT-O	0.86375	0.661125	0.648683	0.654845	0.7859
3	DT-H	0.87425	0.673052	0.680879	0.676943	0.8003
3	LR-O	0.8835	0.777589	0.577554	0.662808	0.9356
3	LR-H	0.89475	0.738386	0.74476	0.741559	0.9435
3	SVM-O	0.89425	0.747326	0.704918	0.725503	0.9391
3	SVM-H	0.89125	0.7	0.806691	0.749568	0.9499

Table 2, Model metrics of one trial

Generally, there are improvements in the performance of a learning model when logistic-regression feature treatment and cost-based outlier removal is incorporated in. The improvements are slight but consistent. Table 3 shows comparison of the average numbers of the model metrics

Model	**Training time (sec)**	**Accuracy**	**Recall**	**Precision**	**F_1 score**	**AUC**	**outlier**
SVM-O	7.8201	0.8986	0.7215	0.7609	0.7404	0.9449	0
SVM-H	7.1821	0.8960	0.8259	0.7058	0.7610	0.9453	372.2
LR-O	0.2523	0.8888	0.6073	0.7914	0.6855	0.9399	0
LR-H	0.2025	0.8970	0.7820	0.7261	0.7526	0.9414	77.6
DT-O	0.3950	0.8593	0.6567	0.6443	0.6503	0.7801	0
DT-H	0.3298	0.8693	0.6725	0.6720	0.6722	0.7862	80.2

Table 3: Average of model evaluation metrics for the hybrid models and original models

All models show a decrease in training time after our hybrid method is employed, this is mostly due to the dimension reduction and outlier removal. We discover that the execution time of the model is notably longer for SVM, with an average of 7.8201 and 7.1821 seconds (benchmark and hybrid, respectively), as opposed to fractions of a second for LR and DT. In addition, while SVM has the highest accuracy among benchmark models at 89.86%, LR has the highest accuracy among the hybrid models at 89.70%. The accuracy for SVM decreases by 0.26% for the hybrid method.

For the other four evaluation metrics (recall, precision, F_1 score, and AUC), DT shows an increase in all but AUC. In fact, DT has consistently the lowest AUC, as SVM and LR have values over 0.9, where DT has values under 0.8; this demonstrated how the performance of DT models are generally poorer. However, DT is the only model which shows an increase in both recall and precision. SVM and LR follow the expected trend of increased recall and decreased precision. LR had the greatest recall increase at 0.1747, indicating that the hybrid LR model is better at predicting the {1}, or default, class. This translates to lower loss for companies, as accurately classifying "risky" borrowers is more important than accurately classifying "good" borrowers. LR shows the greatest improvement in performance, as the difference in AUC is +0.2559.

CONCLUSION AND FUTURE RESEARCH

Based on the results of the research, our hybrid method of logistic-regression-based feature treatment, cost-based outlier removal algorithm combined with SVM, DT, LR model are able to reduces variable dimension, detect and remove outliers with improved performance. The methodology provides a generic approach to refine machine learning algorithm and can be extended to other learning models. This study also provides an evidence that a learning model may work well on certain type of variables but not as effectively on other variables. Hybrid models combine the power of different learning models and are superior in dealing with dataset with complicate feature selection.

However, since we only reduce a small fraction of variables and only test on hybrid model consisting of LR SVM and DT, the performance improvement in this study are very small. It is believed that more improvements can be achieved by reducing more variables, different combinations of traditional learning models. Our data is from a public source, and may not be as comprehensive as data collected by private credit companies. In the future, more research can be conducted on neural networks, In addition, other outlier removal techniques such as wavelet-based outlier removal will be explored as well.

ACKNOWLEDGEMENTS

We would like to thank Dr. Xiaodi Wang from Western Connecticut State University for his help and support throughout our research process. As a professor and mentor, Dr. Wang helped us familiarize ourselves with different machine learning techniques, as well as the classification problem of credit modelling. In addition, he helped provide suggestions on creating our model,

guiding us into different directions that we could explore for new ideas. Without his help and mentorship, our research would not be the same. We are very thankful for his support.

REFERENCES

[1] Baesens, B., Gestel, T. V., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, *54*(6), 627–635. doi: 10.1057/palgrave.jors.2601545

[2] Chen, F.-L., & Li, F.-C. (2010). Combination of feature selection approaches with SVM in credit scoring. *Expert Systems with Applications*, *37*(7), 4902–4909. doi: 10.1016/j.eswa.2009.12.025

[3] Desai, V. S., Crook, J. N., & Overstreet, G. A. (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, *95*(1), 24–37. doi: 10.1016/0377-2217(95)00246-4

[4] Frohlich, H., Chapelle, O., & Scholkopf, B. (2003). Feature selection for support vector machines by means of genetic algorithm. *Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence*. doi: 10.1109/tai.2003.1250182

[5] Gandhi, R. (2018, July 5). Support Vector Machine - Introduction to Machine Learning Algorithms. Retrieved from https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[6] Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, *33*(4), 847–856. doi: 10.1016/j.eswa.2006.07.007

[7] Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767–2787. doi: 10.1016/j.jbankfin.2010.06.001

[8] Lee, T.-S., Chiu, C.-C., Lu, C.-J., & Chen, I.-F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications*, *23*(3), 245–254. doi: 10.1016/s0957-4174(02)00044-1

[9] Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*, *21*(2), 117–134. doi: 10.1016/j.sorms.2016.10.001

[10] (2013, January 5). Peer review. Retrieved from https://www.economist.com/schumpeter/2013/01/05/peer-review

[11] Somol, P., Baesens, B., Pudil, P., & Vanthienen, J. (2005). Filter- versus wrapper-based feature selection for credit scoring. *International Journal of Intelligent Systems*, *20*(10), 985–999. doi: 10.1002/int.20103

[12] Soni, D. (2019, July 16). Dealing with Imbalanced Classes in Machine Learning. Retrieved from https://towardsdatascience.com/dealing-with-imbalanced-classes-in-machine-learning-d43d6fa19d2

[13] Tsai, C.-F., & Chen, M.-L. (2010). Credit rating by hybrid machine learning techniques. *Applied Soft Computing*, *10*(2), 374–380. doi: 10.1016/j.asoc.2009.08.003

[14] Waad, B., Ghazi, B. M., & Mohamed, L. (2013). A Three-Stage Feature Selection Using Quadratic Programming For Credit Scoring. *Applied Artificial Intelligence*, *27*(8), 721–742. doi: 10.1080/08839514.2013.823327

[15] West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, *27*(11-12), 1131–1152. doi: 10.1016/s0305-0548(99)00149-5

[16] Yobas, M. B. (2000). Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics*, *11*(2), 111–125. doi: 10.1093/imaman/11.2.111

DATA REFERENCE

[17] Lending Club's Loan Origination Data, https://www.lendingclub.com/auth/login?login_url=%2Finfo%2Fdownload-data.action