Housing Loan Default Risk Analysis
Using Machine Learning

Soh Jing Guan

*Corresponding author. Email: zingguan9778@gmail.com

ABSTRACT

Housing loan is one of the risky loan for banking and lending institutions due to high demand and it usually involved huge amount of funds. In this research, to minimize the loss of banking and lending institutions due to housing loan default, housing loan default classification model been configured. Four machine learning algorithms including logistic regression, support vector machine, decision tree and random forest been compared and evaluated which algorithm perform the best in housing loan default classification model based on their precision, recall and f1-score. At the same time, this research also focusses on handling data imbalanced problem which could limit the performance of models in detecting positive cases. Data imbalanced is a data issue that always faced in loan default related tasks. In this research, Synthetic Minority Oversampling Technique (SMOTE) been proposed to handle data imbalanced problem. As a result, among four machine learning algorithms, support vector machine achieved best result which is 67% of precision, 90% of recall and 77% of f1-scores. Meanwhile, SMOTE method have been proved enhance the model performance in detect positive cases since each of the model's recall increase after applying SMOTE.

Keywords: *Machine Learning, Loan Default Classification, Resampling, SMOTE*

1. INTRODUCTION

Loan interest fee is the core income for banking and lending institutions. [1] Borrowers need to pay interest fee every month if applied loan from banking and lending institutions. However, loan default may happen and lead those institutions lose their funds and profits.

Loan default is not rare, for China, 9 million loan defaulters have been detected in 2018. For United States, millions of student loans go into default every year. [2]

Among various loans, housing loan is one of the risky loan because it involved huge amount of money. According to HousingWatch, housing loan generally will offer up to 90% of the house values [3]. In this situation, once house loan default happens, related institutions need to afford big amount of lose. Therefore, analysis for housing loan default is crucial to minimize loss due to default.

Machine learning is a subset of artificial intelligence. It is focus to imitate way that human learn by using data and algorithms. Machine learning able to learn from historical data and make prediction to future data. It is popular in financial and healthcare filed to assist decision making. To assists banking and lending institutions in preventing housing loan default. Machine learning algorithm will be used to classify housing loan defaulter.

According to James Lee, there are a lot of machine learning algorithms available, each of them have unique features and may not suite in all cases [4]. To decide which algorithm to use in specific use case, comparison and evaluation between several algorithms is needed. Four machine learning algorithms will be used in this research which are logistic regression, support vector machine, decision tree and random forest, comparison between their performance will be made to evaluate which algorithm most suitable in housing loan default classification.

The performance of machine learning model is very depending on the quality of input data. There is a theorem regarding machine learning which is "rubbish in, rubbish out". If the data input for machine learning training process is dirty. The result of machine learning model will be limited. Therefore, data pre-processing is the most time consuming and important process in machine learning related tasks. However, for loan default related research, data imbalanced is a challenge that always exits. Data imbalanced mean the number of different target classes in a dataset has a huge different.

According to the public resource website Google Developer, if proportion of minority class is below 40%, will treat as data imbalanced. [5] Machine learning may not able to identify minority class properly due to few data available only. Generally, in a classification task, minority class is always positive cases. Therefore, recall of model will be limited due to imbalanced problem. Recall is an evaluation metrics to determine how many positives cases been identified correctly. It is very important for loan default classification task since lending and banking institutions would like to minimize the defaulter as good as possible since huge amount of lose will happen if loan default exits. There are lack of research related handling data imbalanced for housing loan default classification. Hence, this research will also investigate to handle data imbalanced to improve recall for housing loan default classification model. Synthetic Minority Oversampling Technique (SMOTE) will be used to dealing with data imbalanced.

1.1. Objectives for this research.

(1) To identify housing loan defaulter by using machine learning classification.

(2) To identify most accurate machine learning algorithms that train with provided housing loan default dataset by compare and evaluate their performance evaluation metrics.

(3) To handle imbalance data by using resampling algorithm

2. RELATED WORK

2.1. Loan Default Classification with Machine Learning

Jose Romulo [6] proposed a study regarding loan default classification models that focus on housing loan and low income families. In his research, he was found bagging algorithms have the highest accuracy which can reduce default rate from 11.80% to 2.95%. Besides that, dataset been divided into different times intervals of default and the performance of machine learning models also performs better as time interval of default increase.

Peter and Dominique [7] conduct a credit risk analysis using machine learning and deep learning. As a result, instead of deep learning models, random forest and gradient boosting outperform in this study. Moreover, the performance of models enhance again once redundant features eliminated by feature selection method.

Bhoomi Patel and Harshal Patil [8] construct a loan classification model using logistic regression, gradient boosting, catboost and random forest. Among each algorithm, catboost perform the best in his research. Meanwhile, after some exploratory data analysis, features such as job stability and occupation have a negative correlation with target feature.

Lin Zhu and Dafeng Qiu [9] proposed a study regarding loan default classification based on random forest algorithm. They applied SMOTE to handle data imbalanced and using Recursive Feature Elimination to remove redundant features. As a result, Random Forest achieve 98% of accuracy.

2.2. Resampling

Talha Mahboob Alam and Kamran Shaukat [10] conduct a study regarding credit card default classification but more focus on data imbalanced issue. They compare the performance of models before and after balanced the dataset. As a result, balanced data as a input for machine learning model able to enhance their performance. Besides that, they also compare two types of resampling techniques which are oversampling and undersampling. After comparison, they found oversampling technique could perform better.

Ahmad Al-Qerem and Ghazi Al-Naymat [11] focus on exploring significant role of resampling in default prediction model. The model performance in classify minority class has significant enhanced after applying SMOTE.

Ya-Qi Chen and Jian Jun Zhang [12] proposed another research that also focus on handling data imbalanced. Hybrid undersampling method (DSUS) been used to handle data imbalanced for loan default classification model. In a nutshell, DSUS perform better than other resampling techniques including random oversampling and random undersampling.

3. APPROACH AND METHODOLOGY

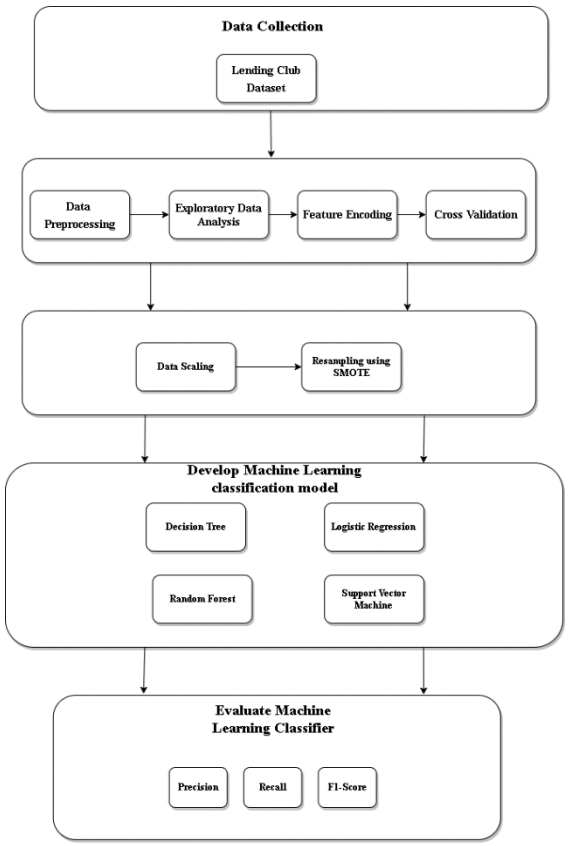

Figure 1 Research Framework

3.1 Data Collection

Dataset used in this research is Lending Club Loan Data from Kaggle. [13] The dataset contains 2260701 of instances and 151 features. Inside this dataset consists information about each applicant, their loan details and status of the loan after approval. This dataset not only contain housing loan data only, also contains other types of loan like Medical loan, Car loan and etc. Since this research focus on housing loan data only, those not related been discarded. Total instances available in this research is 14136. Meanwhile, this dataset faces data imbalanced issues. Proportion for default label is 33% and 67% for non-default label.

3.2 Data Pre-processing

The target feature in this dataset is "loan_status". It contains status of the approved loan. In this dataset, there are 7 categories in the target feature, which been stated in table 1.

Table 1. Different categories in target features of Lending Club Dataset

Categories	Meaning
Default	Based on Lending Club policy, borrower overdue 120 days or more been defined as default.
Fully Paid	Loan already been fully paid by borrower
Charged Off	No expectation for further payment from borrower. Status change to "Charged Off" after 30 days of "Default"
In Grace Period	Late payment but still below 16 days
Late (16-30 days)	Borrower overdue 16 to 30 days
Late (31-120 days)	Borrower overdue 31 to 120 days
Current	Loan is up to date on all outstanding payments

Based on Basel II [14], a loan been categories as default when loan is past due more than 90 days. Therefore, "Default" and "Charged off" will treat as default label. "Fully Paid" treat as non-default label.

```
0    5666
1    1588
Name: loan_status, dtype: int64
```

Figure 2 Number of instances of default and non-default categories.

Based on figure 2, after discard those irrelevant categories in target feature, total amount of instances available is 7254. Ratio of default category become 21% and non-default as 79%.

For missing values in the dataset, those features are empty will discard directly. For those features contains 30% and above of missing values will be discard if don't have above 50% or below -50% of correlation with target features.

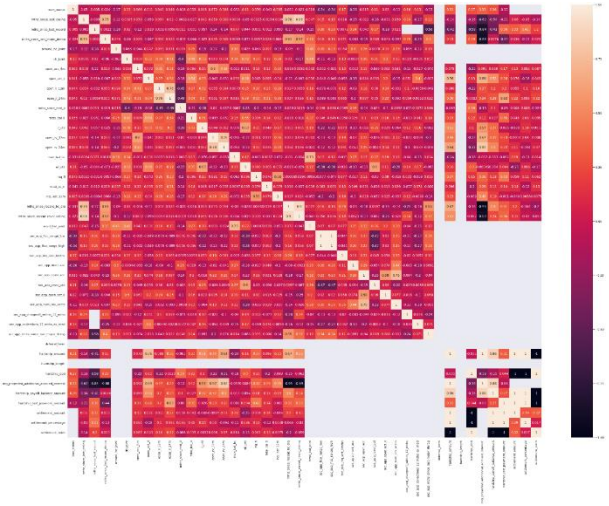

Figure 3 Correlation Heatmap for features contains 30% of missing values and target feature.

As we can see in figure, don't have any feature correlated with our target feature 'loan_status'. So they will be discard.

For other remaining features, median imputation been used to fill in missing values for numerical features. "Unknown" use to fill in missing values for categorical features.

Furthermore, when performing classification in real cases. We only can able use that information that available during loan applying process to classify does the applicant is defaulter or not. So to make sure our model suit in real cases. We need to remove those leaked data, which not available before the loan been approved. Those leaked data will be identified based on data dictionary [15] and help desk that provided by Lending Club. An article write by Genna [16] also provide some support in identify leaked data.

Lastly, outliers in the dataset will be discard based on interquartile range.

3.3 Exploratory Data Analysis.

During exploratory data analysis, some of the features been found only contains one unique value or most of the value are the same, those features will be discard because model not able to explore pattern in these kind of features. Those features including "pymnt_plan", "application_type","disbursement_method", "hardship_flag", "policy_code" and "emp_title". Besides that, "url" and "id" been removed because values in these features are unique for each applicant. Lastly, "grade" been removed because it is a duplicate feature.

At the same times, some of the features may important to model been identified.

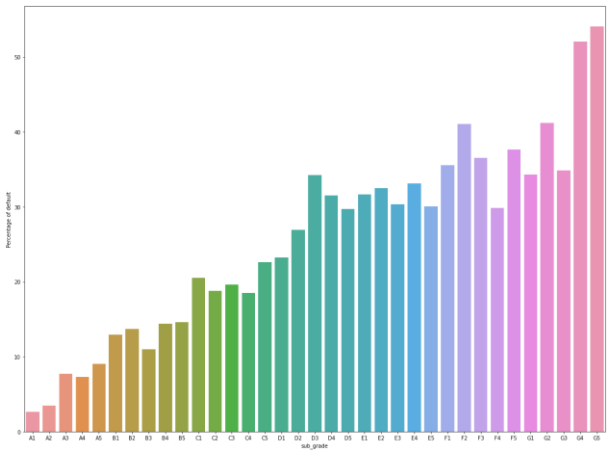

Figure 4 Percentage of default for different loan grades.

The loan grades are evaluating by Lending Club based on those applicant's trustworthiness. As we can see, those applicants been classifying as low grades have higher possibility to default.

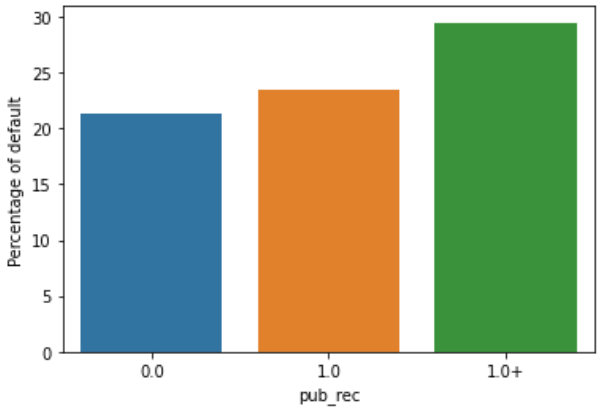

Figure 5 Percentage of default in different times of derogatory public record.

Figure 4 show that percentage of default increased as number of times the applicant involved in derogatory public record. This could be because those related applicants usually don't have much responsibility for others.

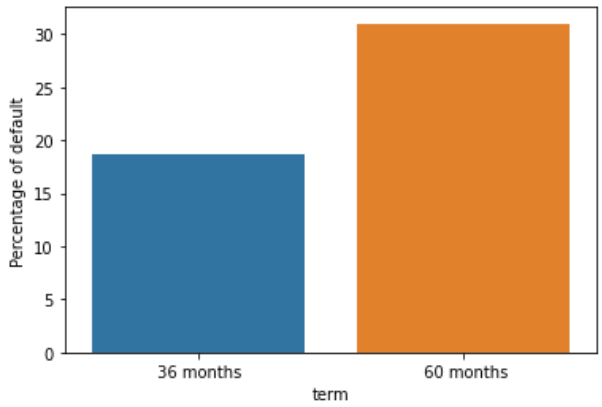

Figure 6 Percentage of default in different loan term

Percentage of default for 60 months' loan term is higher than those apply for 36 months' loan term. This may because when the loan term longer, applicant money flow will be constraints by interest fee longer time which lead to less saving and need proper fund management. If any problem exists in their money flow again, for example like emergency, their available funds will be more limited.

Table 2 will show those remaining features that will process with model training and testing.

Table 2. Features for model training and testing

Features	Description
loan status	Current status of the loan
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
loan_amnt	Amount of loan apply by the applicant
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations
home ownership	The home ownership status provided by the borrower during registration.
verification status	Verified income (whether or not pay slips or a bank statement have been verified by the Lending Club) values are verified, not
annual_inc	verified, and source verified
emp_length	Self-reported annual income provided by the borrower during registration.
num_bc_tl	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
num_il_tl	
pub_rec_bankruptcies	Number of bankcard accounts
revol_util	Number of installment accounts
tax_liens	Number of public record bankruptcies
funded_amnt	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
funded_amnt_inv	Number of tax liens
avg_cur_bal	The total amount committed to that loan
tot_cur_bal	The total amount committed by investors for that loan
installment	Average current balance of all accounts
total_acc	Total current balance of all accounts
tot_coll_amt	The monthly payment owed by the borrower if the loan originates.
Delinquencies	The total number of credit lines currently in
pub_rec	the borrower's credit file
open_acc	Total collection amounts ever owed
acc_delinq	The number of delinquencies in the
revol_bal	borrower's credit file for the past 2 years
sub_grade	Number of derogatory

	public records
tot_hi_cred_lim	The number of open credit lines in the borrower's credit file.
total_bc_limit	The number of accounts on which the borrower is delinquent.
total_il_high_credit_limit	Total credit revolving balance
int_rate	Lending Club assigned loan subgrade
fico_score	Total high credit/credit limit
addr_state	Total bankcard high credit/credit limit
mort_acc	Total installment high credit/credit limit
bc_util	Interest Rate on the loan
earliest_cr_line	Borrower's FICO
initial_list_status	The state provided by the borrower in the loan application

3.4 Feature Encoding

Feature encoding is a process to transfer categorical features to numerical features. [17] Most of the machine learning algorithms not able to handle categorical features, so feature encoding is required to transform those feature to numerical features. Label encoder been selected to conduct feature encoding for this research. It will transform categorical values in a feature to numbers and those numbers in the feature will be interpret by machine learning model that those numbers have some kind of order. [18]. So, it appropriate when categorical features are ordinal

3.5 Cross Validation

Stratified Cross Validation been used for cross validation. It is a variation of k-folds cross validation. The different between it and k-folds cross validation is it make sure proportion of different target classes in each folds are same. It is suitable for imbalanced data problem since it makes sure majority class wont over present in each folds during model training. Evaluate model evaluation with k-folds cross validation also able to evaluate the generalization of the models. 10 folds' cross validation been selected for this research.

3.6 Data Scaling

Data scaling is a technique to transform range of values in each features to same scale. It is important because some of the algorithms like support vector machine and logistic regression need to calculate distance between data, if distribution of each features has huge different will decrease those models performance. [19] StandardScaler been used in this study to perform data scaling. It will transform data into mean value 0 and standard deviation of 1. It is useful for those algorithms perform distance measurement and weight input. [20] Data scaling will perform in each folds, before model training.

$$z = \frac{(x-\mu)}{\sigma} \quad (1)$$

Equation (1) show the formula of StandardScaler. A value in a feature represent as x, μ is mean of the feature, σ is the standardiviation of the feature. Once x subtracts with μ and divide with σ will come out with value been standardized which is z. [21]

3.7 Resampling

SMOTE has been selected to perform resampling in this research. SMOTE is an oversampling method. It increases the amount of minority class until equal to majority class by generating synthetic samples using k-nearest neighbour (KNN).

$$x_{new} = xi + (x_i' - xi) * \alpha \quad (2)$$

The x_new represent synthetic value been generated by SMOTE. x_i^' is belong to one of the KNN for xi. For $\alpha \in [0,1]$ represent real random number. [22]

For k_neighbors of SMOTE method will set to 5 and sampling_technique set to "minority"

3.8 Develop Machine Learning Model

3.8.1 Logistic Regression (LR)

Logistic regression is a machine learning algorithm that suitable for binary classification. Logistic regression weighted input and passed into sigmoid activation function. If output above 0.5 will classify as 1 and classify to 0 of below 0 [23]

$$S(z) = \frac{1}{1+e^{-z}} \quad (3)$$

S (z) is the probability estimated. Once value of S (z) above decision boundary (normally 0.5) the output will be classifying to 1 and if below decision boundary will classify to 0. z is the input to the function and e is the base of natural log

3.8.2 Support Vector Machine (SVM)

SVM is suitable for classification and regression problem. In create a hyperplane to separate data info different classes.

$$K(xi, xj) = \Phi(xi) \cdot \Phi(xj) \quad (4)$$

How a hyperplane been generated is based on the kernel function set for Support Vector Machine, which is K(xi,xj). It is based on the input vectors of xi xj. [24]

In this research, kernel for Support Vector Machine is linear.

3.8.3 Decision Tree (DT)

Decision Tree tries to form conditions on the features to separate all the classes in the dataset. Decision tree selects features as conditions to split the dataset into different group based on information gain.

$$Gain\ (T, X) = Entropy(T) - Entropy(T, X) \quad (5)$$

3.8.4 Random Forest (RF)

Random Forest establish multiple decision trees and use the result that gets by majority trees.

$$\sum_{i=1}^{C} -f_i \log(f_i) \quad (6)$$

	Logistic Regression	Support Vector Machine	Decision Tree	Random Forest
Fast	✓			
Not Sensitive to outliers			✓	✓
Suitable for small dataset	✓	✓		
Missing Value less influence			✓	✓
Low probability of overfitting	✓			✓
No need Data Scaling			✓	✓

Figure 7 Comparison between each Machine Learning algorithms

3.9 Model Evaluation

The performance of each models will evaluate based on average precision, recall and f1-score in 10 folds.

3.9.1 Precision

Precision calculate percentage of the model's results that are correct and relevant.

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (7)$$

3.9.2 Recall

Recall calculating percentage of total relevant results correctly classified by the algorithms.

$$recall = \frac{true\ positive}{true\ positive + false\ negatives} \quad (8)$$

3.9.3 F1-score

F1-score consider both precision and recall. It determines the balanced between precision and recall.

$$F_1 = 2 * \frac{precision * recall}{precision + recall} \quad (9)$$

3.10 Baseline Model

Although we could found some researches using same dataset with this research. But they using full data and this research only extract data regarding housing loan. Since dataset for this research and other not exactly the same, we can't find benchmark model to

compare. Based on Aslam [25], decision tree is a popular classification algorithm in credit scoring domain. It also eases to configure. Therefore, this research will use decision tree as baseline model to compare with other three models.

4. RESULT

Machine Learning Model	Precision	Recall	F1-Score
Logistic Regression	0.67	0.88	0.76
Support Vector Machine	0.67	0.90	0.77
Decision Tree	0.62	0.70	0.66
Random Forest	0.67	0.83	0.74

Figure 8 Average performance of each models based on 10 folds.

Based on figure 8, support vector machine gets the highest f1-score and recall which is 77% and 90%. In the same time, baseline model decision tree only gets 66% of f1-score and 70% of recall. The best model in this research is better than the baseline model

Table 3. Precision of models in each folds

Folds	LR	SVM	DT	RF
1	0.67	0.67	0.62	0.67
2	0.67	0.67	0.61	0.67
3	0.68	0.67	0.62	0.69
4	0.68	0.67	0.63	0.67
5	0.67	0.66	0.62	0.68
6	0.67	0.67	0.62	0.66
7	0.67	0.65	0.62	0.67
8	0.67	0.67	0.62	0.68
9	0.69	0.67	0.62	0.67
10	0.67	0.67	0.62	0.67

Table 4. Recall of models in each folds

Folds	LR	SVM	DT	RF
1	0.87	0.90	0.70	0.84
2	0.88	0.89	0.68	0.82
3	0.88	0.90	0.70	0.86
4	0.87	0.90	0.69	0.83
5	0.88	0.89	0.70	0.84
6	0.88	0.90	0.70	0.83
7	0.88	0.90	0.70	0.82
8	0.86	0.88	0.71	0.83
9	0.88	0.90	0.70	0.82
10	0.88	0.90	0.70	0.83

Table 5. Precision of models in each folds

Folds	LR	SVM	DT	RF
1	0.76	0.77	0.66	0.74
2	0.73	0.77	0.66	0.74
3	0.76	0.77	0.66	0.76
4	0.76	0.76	0.67	0.74
5	0.75	0.77	0.66	0.75
6	0.76	0.77	0.66	0.74
7	0.76	0.78	0.68	0.74
8	0.75	0.77	0.66	0.75
9	0.76	0.78	0.65	0.74
10	0.76	0.77	0.66	0.74

Based on table 3, 4 and 5, we could found performance in each folds are very similar, only minor different. This has indicated that our model didn't bias to certain data, they can perform consistently.

	Training			Testing		
Machine Learning Model	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Logistic Regression	0.88	0.91	0.89	0.67	0.88	0.76
Support Vector Machine	0.88	0.91	0.89	0.67	0.90	0.77
Decision Tree	1.0	1.0	1.0	0.62	0.70	0.66
Random Forest	0.99	0.99	0.99	0.67	0.83	0.74

Figure 9 Average Model Performance for training set and testing set based on 10 folds

However, figure 9 show the performance of models in training set and testing set. Decision tree and random forest may face overfitting problem due to huge different result in training set and testing set. This may because these two algorithm are complex algorithm and the size of the dataset in this research is not enough for them.

	Without SMOTE			With SMOTE		
Machine Learning Model	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Logistic Regression	0.74	0.74	0.74	0.67	0.88	0.76
Support Vector Machine	0.73	0.74	0.73	0.67	0.90	0.77
Decision Tree	0.64	0.66	0.65	0.62	0.70	0.66
Random Forest	0.74	0.73	0.73	0.67	0.83	0.74

Figure 10 Machine Learning models before and after SMOTE

Lastly, the after applying SMOTE method to handling data imbalanced problem, each model's precision are slightly decrease but f1-score increase slightly and recall is significantly increase. Especially for support vector machine, its recall increased from 74% to 90% which is total 19% of incensement. As recall for each models increased, which indicated each models perform better in detecting positives cases or minority cases. Therefore, with handling data imbalanced using SMOTE, each machine learning models perform better in detecting defaulter.

5. CONCLUSION

In conclusion, support vector machine performs the best in this research, which gets 67% of precision, 90% of recall and 77% of f1-scores. Besides that, after applying SMOTE, each models perform better in detecting defaulter.

In future work, hyperparameter tuning could be consider to decrease complexity of decision tree and random forest in order to avoid overfitting. At the same time, by implement hyperparameter tuning, performance of support vector machine and logistic regression also able to maximize. Besides that, other resampling techniques could be considering in next analysis to identify best resampling technique to handling data imbalanced.

Lastly, one of the challenge for housing loan default classification is lack of dataset available. Meanwhile, the lending club dataset has a lot of data problem like missing values and leaked data which laid data pre-processing become more time-consuming. Also, the size of the house loan in Lending Club dataset is below 10 thousand which may be too small to conduct more powerful algorithms like deep learning algorithms. To further improve housing loan default analysis, additional dataset is required.

REFERENCES

[1] Sang, A. (2021). How Do Banks Make Money? https://www.clevergirlfinance.com/blog/how-do-banks-make-money/

[2] Shoumo, S. Z. H., Dhruba, M. I. M., Hossain, S., Ghani, N. H., Arif, H., & Islam, S. (2019). Application of Machine Learning in Credit Risk Assessment: A Prelude to Smart Banking. IEEE Region 10 Annual International Conference, Proceedings/TENCON, 2019-Octob, 2023–2028. https://doi.org/10.1109/TENCON.2019.8929527

[3] Consumer Guide on Getting a Housing Loan. (2021).https://www.housingwatch.my/05_02_guide_housingloan.html

[4] James Le. (2021). The Top 10 Machine Learning Algorithms Every Beginner Should Know. https://builtin.com/data-science/tour-top-10-algorithms-machine-learning-newbies

[5] Developer, G. (n.d.). Data Preparation and Feature Engineering for Machine Learning. https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data

[6] de Castro Vieira, J. R., Barboza, F., Sobreiro, V. A., & Kimura, H. (2019). Machine learning models for credit analysis improvements: Predicting low-income families' default. Applied Soft Computing Journal, 83, 105640. https://doi.org/10.1016/j.asoc.2019.105640

[7] Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep learning models. Risks, 6(2), 1–20. https://doi.org/10.3390/risks6020038

[8] Patel, B., Patil, H., Hembram, J., & Jaswal, S. (2020). Loan default forecasting using data mining. 2020 International Conference for Emerging Technology, INCET 2020, 9–12.

[9] Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. Procedia Computer Science, 162(Itqm 2019), 503–513. https://doi.org/10.1016/j.procs.2019.12.017

[10] Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., Li, J., & Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets. IEEE Access, 8, 201173–201198. https://doi.org/10.1109/ACCESS.2020.3033784

[11] Al-Qerem, A., Al-Naymat, G., Alhasan, M., & Al-Debei, M. (2020). Default prediction model: The significant role of data engineering in the quality of outcomes. International Arab Journal of Information Technology, 17(4 Special Issue), 635–644. https://doi.org/10.34028/iajit/17/4A/8

[12] Chen, Y. Q., Zhang, J., & Ng, W. W. Y. (2018). Loan Default Prediction Using Diversified Sensitivity Undersampling. Proceedings - International Conference on Machine Learning and Cybernetics, 1, 263–268. https://doi.org/10.1109/ICMLC.2018.8526936

[13] Kaggle. (2019). All Lending Club loan data. https://www.kaggle.com/wordsforthewise/lending-club

[14] Bank Negara Malaysia. (2019). Capital Adequacy Framework (Basel II - Risk Weighted Assets). Bank Negara Malaysia, June, 1–497.

[15] Lending Club. (n.d.-a). Lending Club Data Dictionary. https://resources.lendingclub.com/LCDataDictionary.xlsx

[16] Edwards, G. (n.d.). Data Cleaning and Preparation for Machine Learning. https://www.uxax.org/post/data-cleaning-and-preparation-for-machine-learning

[17] Sarangpratap. (2021). Feature Encoding Techniques –Machine Learning. https://www.geeksforgeeks.org/feature-encoding-techniques-machine-learning/

[18] Brownlee, J. (2017). Why One-Hot Encode Data in Machine Learning? https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

[19] VERMA, Y. (2021). Why Data Scaling is important in Machine Learning & How to effectively do it. https://analyticsindiamag.com/why-data-scaling-is-important-in-machine-learning-how-to-effectively-do-it/

[20] Jaadi, Z. (2021). When and Why to Standardize Your Data? buildin. https://builtin.com/data-science/when-and-why-standardize-your-data

[21] Shuai, Y., Jiang, C., Su, X., Yuan, C., & Huang, X. (2020). A Hybrid Clustering Model for Analyzing COVID-19 National Prevention and Control Strategy. 2020 IEEE 6th International Conference on Control Science and Systems Engineering, ICCSSE 2020, 68–71. https://doi.org/10.1109/ICCSSE50399.2020.9171941

[22] Alam, T. M., Shaukat, K., Hameed, I. A., Luo, S., Sarwar, M. U., Shabbir, S., Li, J., & Khushi, M. (2020). An investigation of credit card default prediction in the imbalanced datasets. IEEE Access, 8, 201173–201198. https://doi.org/10.1109/ACCESS.2020.3033784

[23] Kanani, B. (2019). Machine Learning Tutorials. https://studymachinelearning.com/logistic-regression/

[24] de Castro Vieira, J. R., Barboza, F., Sobreiro, V. A., & Kimura, H. (2019). Machine learning models for credit analysis improvements: Predicting low-income families' default. Applied Soft Computing Journal, 83, 105640. https://doi.org/10.1016/j.asoc.2019.105640

[25] Aslam, U., Aziz, H. I. T., Sohail, A., & Batcha, N. K. (2019). An empirical study on loan default prediction models. Journal of Computational and Theoretical Nanoscience, 16(8), 3483–3488. https://doi.org/10.1166/jctn.2019.8312