Volume:6 Issues: 32 [March, 2021] pp. 1 - 14]
International Journal of Accounting, Finance and Business (IJAFB)
eISSN: 0128-1844
Journal website: www.ijafb.com

SOCIAL MEDIA DATA TO IMPROVE CREDIT SCORING ACCURACY WITH A DATA MINING APPROACH BASED ON SUPPORT VECTOR MACHINE: CASE STUDY OF AN ONLINE PEER TO PEER LENDING IN INDONESIA

Okta Saputra[1]
Taufik Faturohman[2]
Sudarso Kaderi Wiryono[3]

[1]School of Business and Management, Institut Teknologi Bandung, Indonesia,
(E-mail: okta_saputra@sbm-itb.ac.id)
[2]School of Business and Management, Institut Teknologi Bandung, Indonesia,
(E-mail: taufik.f@sbm-itb.ac.id)
[3]School of Business and Management, Institut Teknologi Bandung, Indonesia,
(E-mail: sudarso_kw@sbm-itb.ac.id)

Article history
Received date : 27-8-2020
Revised date : 28-8-2020
Accepted date : 9-3-2021
Published date : 31-3-2021

To cite this document:
Saputra, O., Faturohman, T., & Kaderi Wiryono, S. (2021). Social Media Data to Improve Credit Scoring Accuracy with A Data Mining Approach Based on Support Vector Machine: Case Study of An Online Peer to Peer Lending in Indonesia. *International Journal of Accounting, Finance and Business (IJAFB)*, 6 (32), 1 - 14.

Abstract: *In recent years, financial technology (fintech) is rapidly expanding in Indonesia. The fintech ecosystem in Indonesia is dominated by Peer to Peer (P2P) lending. Small and micro-enterprises and individual borrowers do not need loan guarantors and collateral in getting the financing. Yet, this condition will position P2P Lending to be exposed with credit risk. The output of data mining technique can be used to make the credit scoring model better, one of the algorithms by using Support Vector Machine. There is one online peer to peer lending company in Indonesia developing social evaluation to enhance their credit scoring to face the fluctuation of non-performing loan. Therefore, the aim of this study is to build credit scoring model using social media data based on Support Vector Machine. The model development process adapted Cross-Industry Standard Process for Data Mining (CRISP-DM), which consists of business understanding, data understanding, data preparation, model development, and evaluation. The borrower's data in the company is used including the borrower's demographic, historical payment data, and social media data. The research that has been done resulted that SVM Linear has the best performance compared to other kernels. By adding social media data, it can increase the performance of the credit scoring model measured in AUC as much as 7%.*

Keywords: *Social Media Data, Credit scoring, Data Mining, Support Vector Machine*

Volume:6 Issues: 32 [March, 2021] pp. 1 - 14]
International Journal of Accounting, Finance and Business (IJAFB)
eISSN: 0128-1844
Journal website: www.ijafb.com

Introduction

In recent years, financial technology (Fintech) is rapidly expanding in Indonesia. The growth of Indonesian Fintech industry benefits from various growth drivers that have supported rapid recent expansion especially from the demographic population that relates to the digital landscape. According to Fintech Singapore in Indonesia Fintech Landscape Report (2018), the number of financial technology transactions is projected to reach US$22.3 billion by the end of 2018, increasing 16.3% annually from 2017.

Fintech offers a broad variety of financial services, including transaction settlement, capital raising, investment management, fundraising and distribution insurance, market support, equity crowdfunding, and other financial support and service activities. As May 2019, the fintech ecosystem in Indonesia is dominated by P2P lending with a proportion of 43% and followed by payment as much as 26% (Asian Development Bank Institute, 2019). Fintech Lending or also called Fintech Peer-to-Peer Lending (P2P Lending) is one of the innovations in the financial services with the use of technology which allows the lenders and borrower to make a transaction without having to meet in person. In December 2019, the total lending accumulation is as much as 81.5 trillion rupiahs, increasing 259.56% from last year. (OJK, 2019).

According to the P2PMarketData in 2020, P2P lending generally operates at lower operating costs than other financial institutions, allowing them to provide favourable conditions for both lenders and borrowers by shifting these lower costs to both supply and demand sides. Small and micro-enterprises and individual borrowers who find it difficult to obtain loans from the bank do not need guarantors and collateral to make it easier for them to obtain financing. As a result, in the best-case scenario, lenders are given favourable returns relative to other savings and investment products, and borrowers are provided access to capital, competitive interest rates, and a quicker process compared to comparable products offered by other financial institutions (P2PMarketData, 2020). Yet, this condition will position the P2P Lending to be exposed with credit risk.

Credit risk is the capacity of a borrower failed to meet its obligations under agreed terms (Basel, 2000). In many literatures, credit risk is usually proxied by non-performing loans (NPLs) which are loans and advances overdue by 90 days or more from the due date. According to OJK in Performance Report (2019), the NPL of Fintech Lending in Indonesia faced an increase from last year becoming 3.51% by November 2019.

To manage the risk, the implementation of credit risk management should be done in a financial institution. Greuning and Iqbal (2007) stated that credit risk management is a structural approach to manage uncertainties through risk assessment, developing strategies to manage it, and mitigation of risk using managerial resources. In conducting risk assessment, there is a technique called credit scoring which allows the financial institution to decide whether to accept or reject the credit applicants. According to Hand & Harley (1997), credit scoring is a term used to describe the formal statistical method which is used to classify the credit application into bad credit or good credit. It is necessary to determine credit scoring of a customer because a slight increase in performance of the customer's classification can increase significant financial returns of the financial institution (Hand & Henley, 1997).

Volume:6 Issues: 32 [March, 2021] pp. 1 - 14]
International Journal of Accounting, Finance and Business (IJAFB)
eISSN: 0128-1844
Journal website: www.ijafb.com

Hence, it is very important to increase the accuracy in determining the creditworthiness of customers. One of the ways to increase accuracy is to analyze the factors that affect the performance of customers in fulfilling their duty. The mainly used factors are the socio demographic factors such as age, gender, job, income, and etc. One way to analyze these factors is by using the data mining techniques. Data mining techniques can be done by utilizing historical data owned by the financial institutions. The data mining process involves identifying problems with the business, identifying data mining goals, retrieving the required databases, and using data mining techniques to analyze data with the ultimate goal of getting important results in making strategic decisions. The output can be used to make the credit scoring model better. There are several machine learnings that can be used to develop the credit scoring model, one of them is Support Vector Machine (SVM). SVM is learning algorithms that analyze data used for classification. When using SVM, two problems are confronted: how to choose the optimal input feature subset for SVM and how to set the best kernel parameters. These two problems are crucial because the feature subset choice influences the appropriate kernel parameters and vice versa (Frohlich & Chapelle, 2003).

There is one of online peer to peer lending in Indonesia established in March 2017 and located in Bandung which faced a fluctuate non-performing loan from May – September 2018. To address the issue, this online peer to peer lending is developing a social media evaluation to enhance predictability of its credit scoring in terms of avoiding bad debt borrowers. Furthermore, this online peer to peer lending is also still using the credit scoring model which has not yet involved the data mining approach whereas this also can help to make the credit scoring model better.

Literature Review

Online Peer to Peer Lending Development

World Bank Group (2018) stated that peer to peer lending is an online platform that offers credit application as a new alternative. According to Bank Indonesia (2015), peer to peer lending has the chance to be advanced in Indonesia because the amount of unbankable for small and medium enterprises (SME) which are around 60% - 70% from a total of 56.4 million SME in 2014. Furthermore, the World Bank (2017) determined that Indonesia contributed 6% of 1.7 billion adults who are unbanked globally or do not have a bank account. There are 164 fintech lending has been registered in OJK per December 2019 with a combination of 152 conventional and 12 Islamic peer to peer lending (OJK, 2019). Consider the opportunity and significant growth, it can be expected that online peer to peer lending would have an encouraging development in Indonesia. Director of Regulation of Licensing and Supervision of Fintech in OJK (2018) said OJK will not cover the risk which appears and exposed in those platforms since OJK is only responsible to supervise the registered lending platform even if it has been registered in OJK. whereas online peer to peer lending is exposed by bad debt risk. According to OJK in Performance Report (2019), the NPL of Fintech Lending in Indonesia faced an increase from last year becoming 3.51% by November 2019.

Credit Risk

Credit risk is the capacity of a bank borrower to fail to meet its obligations under agreed terms. Meanwhile, credit risk management is a structured approach to manage ambivalence through risk appraisal, develop methods to manage it, and mitigation of risk using managerial resources

Volume:6 Issues: 32 [March, 2021] pp. 1 - 14]
International Journal of Accounting, Finance and Business (IJAFB)
eISSN: 0128-1844
Journal website: www.ijafb.com

(Greuning & Iqbal, 2007). In general, Ali Fatemi (2016), on his research in Credit Risk Management: A survey of Practices, divided credit risk management into two categories; proprietary (internal) credit risk management, and the vendor-marketed models where companies use the third party that is credit rating agencies

Credit Scoring

Credit scoring is essentially a way of recognizing the different groups in a population when one cannot see the characteristic that separates the groups but only related ones. Chen, Li, & Zeng (2018), on some previous studies about credit scoring, explained that characteristics of good mechanism credit scoring can predict repayment performance of funded loans reflects the positive intermediary role played by the platform. Therefore, the credit scoring model is fundamental to predict the borrower's loan default probability, so that financial institutions could take an optimal decision to avoid any losses coming from credit risk exposure.

Social Credit Scoring

The exploration of customer creditworthiness based on social network information and seeing consumer social status more broadly is used to determine the social credit score (Dellarocas, et.al, 2016). Furthermore, Tan & Phan (2016) in their paper titled Social Media-Driven Credit Scoring: The Predictive Value of Social Structures proved that added social network data could increase predictability rate by 18%. Social tier as well as personal interaction in the borrower's social account can be used to define the borrower's social network and status. The social credit scoring is predicted to make a better individual loans scoring.

Data Mining

Data mining is a technique for extracting patterns or rules from large databases. Malhotra and Kanika (2014) argued that data mining is related to the discovery of a knowledge of data that is useful to be used in the future. There are six standard stages in conducting the data mining process which are business (organizational) understanding, data understanding, data preparation, modelling, evaluation, and deployment (Chalaris, Gritzalis, Maragoudakis, Sgouropoulou, & Tsolakidis, 2014). Here, Construction of credit scoring models requires data mining technique (Bhatia, 2017)

Machine Learning

Machine Learning (ML) a scientific discipline that includes design and development of algorithms that allow computers to develop behaviour that is based on empirical data. Machine Learning can also be interpreted as a discipline that assigns computers to study and act like humans, and improve their learning ability over time automatically, by supplying data and information as forms real world experiences and interactions. Various machine learning classifiers are well-known in research where significant technological impact on society has been observed. Machine learning classifiers are effective in a wide range of business and societal applications (Cubric, 2020). Financial institutions have recently been interested in using machine learning to make predictions, such as credit scoring (Trivedi, 2020)

Support Vector Machine

Support Vector Machine is one of the most performant off-the-shelf supervised machine learning algorithms because it often generates in pretty good results without many tweaks (Kowalczyk, 2017). Supporting vector machines analyse data used for classification and

Volume:6 Issues: 32 [March, 2021] pp. 1 - 14]
International Journal of Accounting, Finance and Business (IJAFB)
eISSN: 0128-1844
Journal website: www.ijafb.com

regression analysis. With a set of examples training, each marked as one or two other categories, SVM algorithms build a model that provides a new example to one category or another category, making it a binary linear classifier non-probabilistic. A support vector engine builds a hyperplane or hyperplane circuits in high-dimensional or unlimited space, which can be used for classification, regression, or other tasks such as outlier detection.

Kernel Tricks

In machine learning, the kernel method is a class of algorithms for pattern analysis, whose most famous members are supporting vector machines (SVM). The general task of pattern analysis is to find and study type general relationships (e.g. group, rank, main component, correlation, classification) in a data set. Kernel method on the use of kernel functions allows them to operate in space based on implicit dimensions, without calculating data coordinates in that space, but only by calculating the inner results among the pictures from all pairs of data in the feature space. This operation is often cheaper rather than explicit calculation of coordinates. This approach is called the "kernel trick". Function The kernel has been introduced for sequence data, graphics, text, images, and vectors. Therefore, the Kernel method is often referred to as a generalized dot product. Assume we have the transformation function $\boldsymbol{\varphi : Rn \rightarrow Rm}$ which aims to do data transformation from $\boldsymbol{Rn}$ headed $\boldsymbol{Rm}$ with m = n +1. (Al-Mejibli et al, 2018)

Methodology

Research Design

This research methodology refers to the stages of Cross Industry Standard Process for Data Mining (CRISP-DM). These stages are adjusted with the current condition. The research design is explained in the figure 1.

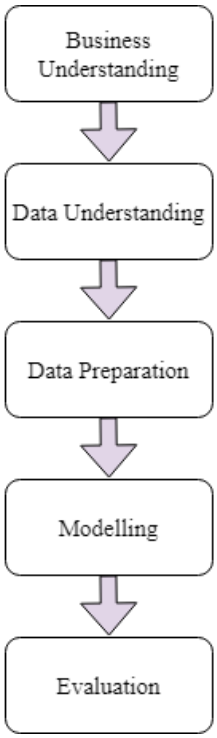

Figure 1: Research Design

Volume:6 Issues: 32 [March, 2021] pp. 1 - 14]
International Journal of Accounting, Finance and Business (IJAFB)
eISSN: 0128-1844
Journal website: www.ijafb.com

Business Understanding

This stage focuses on understanding the main objectives of the data mining project and needs from the company perspective. This first stage will be divided into two parts which are problem identification and literature review.

Data Understanding

At the data understanding stage, the borrower's data is collected and observed. The data that has been collected is studied because understanding good data can help make it easier to determine the data preprocessing techniques.

Data Collection

The borrower's data in the company will be used as research data, including the borrower's demographic, historical payment data, and social media investigation data. These data are confidential. Nonetheless, the company only has the social media account of the borrowers, not the detailed data, so it needs to be investigated manually. The research will be focused on the borrower's information from March 2017 – January 2019. This time frame is only limited for one year and ten months because the company was just built-in 2017. In the meantime, the social media investigation has no time frame because the limitation only focuses on the number of posts in each borrower's social media account.

Data Description

The current company's dataset is comprised with both demographic and historical payment data of the borrowers. Demographic data consists of gender, marital status, district, employment, and income while historical payment data consists of instalment and tenor. These data are called as features (independent variable) and will be used to determine the target variable (dependent variable) which is late. This target variable will categorize into two label which is non-default and default. Referring to the previous research conducted by Tan & Phan (2016), Hsieh, et al. (2011), and Guo, et al. (2016), this research will also use the social media data as additional features. This social media data will be focused on Instagram and Facebook since the company focused to investigate these. The social media data that will be used consists of IG_Month, Posting_Midnight, Followers, Following, IGPost_Month (Guo, et al., 2016) and Rel_Account which refers to the number of religion account followed or affiliated (Tan & Phan, 2011). By total, there are 13 features comprised of demographic, historical payment, and social media data, and one target variable. The explanation of each features is served in the table 2.

Table 2: Features and Target Variables

No	Feature	Data Type	Description
Demographic			
1	GENDER	Categorical	Gender of the borrower consist of 2 categories: MALE and FEMALE
2	MARITAL	Categorical	Marital Status of the borrower consist of 2 categories: SINGLE and MARRIED
3	DISTRICT	Categorical	City where the borrower live, consists of 3 categories: KOTA BANDUNG, KABUPATEN BANDUNG, and OTHER

Volume:6 Issues: 32 [March, 2021] pp. 1 - 14]
International Journal of Accounting, Finance and Business (IJAFB)
eISSN: 0128-1844
Journal website: www.ijafb.com

4	EMPLOYMENT	Categorical	Job of the borrower consist of 4 categories: MANAJER, KARYAWAN, WIRASWASTA, and OTHER
5	INCOME	Float	Monthly income of the borrower (in Rupiah)
Historical Payment			
6	INSTALLMENT	Float	The amount of money that has to be paid as loan (in Rupiah)
7	TENOR	Integer	Duration of the credit (in month)
Social Media			
8	IG_MONTH	Integer	The user's duration using Instagram according to the date of their first posting in form of month (in month)
9	POSTING_MIDNIGHT	Integer	The number of Facebook posts in midnight (00.00-03.59) from the last 21 posts of each account. (in unit)
10	REL_ACCOUNT	Integer	The number of religion account that is followed by borrower in Instagram (in unit)
11	FOLLOWERS	Integer	Number of followers in Instagram (in unit)
12	FOLLOWING	Integer	Number of following in Instagram (in unit)
13	IGPOST_MONTH	Integer	The proportion of the total number of posts divided by months duration using Instagram (in unit)
Target Variable (Dependent Variable)			
14	LATE	Integer	Number of days counted as late in paying the credit (in day)

Analysis and Discussion

In this chapter, there are 3 stages that are processed including data preparation, modelling, and evaluation. The first stage would be data preparation where the data is pre-processed before it is used to develop the model. Here, it is conducted a feature selection to reduce the dimension of features. Next stage is modelling where there are three different model developed from three different kernel of Support Vector Machine (SVM) using the data with selected features from data preparation. This model is firstly developed in baseline and will be optimized into hypertuning parameter. Later, these models will be assessed in evaluation stage to see the result of the optimization. Next, there is only one model of SVM that will be selected after the kernel selection evaluation and will be implemented into three different scenarios with the different numbers of features used. Lastly, this scenario will be evaluated in order to select best scenario which can be suggested to be implemented as company's credit scoring model.

Volume:6 Issues: 32 [March, 2021] pp. 1 - 14]
International Journal of Accounting, Finance and Business (IJAFB)
eISSN: 0128-1844
Journal website: www.ijafb.com

Data Preparation

For this research, the data preparation stage includes data transformation and feature selection. This stage is based on the results of understanding the data in the previous stage. This is done in order to produce data that is ready to be processed.

Data Integration

In this stage, the data is divided into data training and data testing with a proportion of 70% and 30% respectively. The training data is used for the construction of the model, while the test data is used for evaluation of the model that has been built.

Data Transformation

At this stage, changes to the data format are done according to the needs of the model that will be built. In this study, data transformation consists of two steps which are:

1. Data Encoding
   This step is aimed to encode the data so that it can be processed for credit scoring modelling. In order to build the model of this research, the data of target variable should be transformed into binary, which means only two classes which are '0' as non-default and '1' as default. Hence, In the target variable which is named as 'LATE', the data that is valued as 0 remains 0, but for the data that is other than 0 is replaced by 1. Next, the researcher transforms several categorical features in the form of string into numbers because the model that will be developed can only process the numerical data. In this research, one hot encoding is used to transform the features of 'GENDER', 'MARITAL', 'DISTRICT', and 'EMPLOYMENT' from the dataset. This feature transformation causes the dataset that is initially having 14 columns added to 21 columns which also includes the column 'LATE'.
2. Feature scaling
   Feature scaling is done to standardize values vary the range of features, and it is useful to help accelerate the learning process and improve the quality of the model. The scaling is done by utilizing the library in Sklearn, StandardScaler. This will transform the data such that its distribution will have a mean value 0 and a standard deviation of 1.

Feature Selection

Since the data has high dimensionality, feature selection should be done in order to make the model become more efficient with only several features used. Correlation heat map is used for the feature selection. In summary, there are 8 features that have high correlation both positive and negative correlation with 'LATE' as the target variables which can be seen from this table:

Table 3: Correlation Value

Features	Correlation Value
TENOR	0.373300
POSTING_MIDNIGHT	0.260929
FOLLOWING	0.229467
IGPOST_MONTH	0.186938
EMPLOYMENT_KARYAWAN	0.175746
IG_MONTH	-0.178479
EMPLOYMENT_OTHER	-0.185256
INCOME	-0.222132

Volume:6 Issues: 32 [March, 2021] pp. 1 - 14]
International Journal of Accounting, Finance and Business (IJAFB)
eISSN: 0128-1844
Journal website: www.ijafb.com

There is no rule for determining what size of correlation that is considered strong, moderate, or weak. The interpretation of the coefficient depends, in part, on the topic of study. Here the researcher will only select those features that have value more than 15% of correlation both for positive and negative value. Hence, the features that are correlated with target variables are TENOR, POSTING_MIDNIGHT, FOLLOWING, IGPOST_MONTH, EMPLOYMENT_KARYAWAN, EMPLOYMENT_WIRASWASTA, IG_MONTH, EMPLOYMENT_OTHER, and INCOME. These features will be used to develop the model in the next stage.

Modelling

At this stage, the classification model is designed and implemented. This stage processes the training data available to create a credit scoring model by using svm.SVC on Scikit-Learn library. The model is divided into two categories which are the baseline and the hypertuning parameter

Baseline

For the baseline, Support Vector Machine (SVM) will be separated into 3 models by using 3 different kernels (Linear, RBF, and Polynomial) in order to see which kernel will perform better for the data set. All of these three kernels parameters will be set into default in order to see the performance point of the models.

Hypertuning Parameter

Hypertuning Parameter is done to get the best parameter values that can be used to optimize each baseline model of the SVM kernels. This is done by implementing an optimization method named Grid Search in which this is an exhaustive search method that conducts the experiment on all combinations of parameters that have been previously defined.

1. **SVM Linear**
   According to Drucker, Wu, & Vapnik (1999), one of the advantages of the linear kernel SVM has only to tune for parameter C. C parameter in SVM is the penalty parameter of the error term. It's considered as the degree of correct classification that the SVM has to meet. For greater values of C, SVM will look for a higher margin and this cause the model can classify perfectly. Yet, increasing C too much has a risk to lose the generalization properties of the classifier. There is no rule of thumb to choose a C value, it totally depends on the training data. Hence, the parameter c is tested with 8 alternatives values which are 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000. From the result, parameter C is found to be 0.1. This parameter is then re-implemented to the SVM Linear algorithm to build the optimized model.
2. **SVM RBF**
   With the RBF kernel, there are two parameters to be determined in the SVM model: C and gamma. The grid search approach (Hsu, Chang, & Lin, 2003) is an alternative to finding the best C and gamma when using the RBF kernel function. Gamma parameter is the inverse of the standard deviation of the RBF kernel (Gaussian function), which is used as a similarity measure between two points. If gamma is too large, the radius of the area of influence of the support vectors only includes the support vector itself and no amount of regularization with C will be able to prevent overfitting. When gamma is very small, the model is too constrained and cannot capture the complexity or "shape" of the data. The combination of parameter that is tested is gamma with 8 alternatives value which are 'auto',

Volume:6 Issues: 32 [March, 2021] pp. 1 - 14]
International Journal of Accounting, Finance and Business (IJAFB)
eISSN: 0128-1844
Journal website: www.ijafb.com

0.0000001, 0.000001, 0.00001, 0.0001, 0.001, 0.01, 0.1 and 8 alternatives of C value which are 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000. From the result, the tuning parameter C is found to be 1 and gamma 'auto' which found to be the same with the baseline model. These parameters are then re-implemented to the SVM RBF algorithm to build the optimized model.

3. **SVM Polynomial**

According to the previous study conducted by Belotti and Crook in 2009, they built the model by using SVM Polynomial. They also tune the parameters which are C and degree in order to get a more accurate model. Degree is a parameter used when the kernel is set to 'poly'. It's basically the degree of the polynomial used to find the hyperplane to split the data. By using Grid Search, the combination of parameters degree with 3 alternatives which are 2, 3, 4 and 8 alternatives of C value which are 00.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000. From the table above, parameter C is found to be 4 with degree 3. These parameters are then re-implemented to the SVM Polynomial algorithm to build the optimized model.

Evaluation

In this stage, the previous three different kernels model with only 8 features that been built will be evaluated in order to see the performance of model before and after the hypertuning parameter optimization using grid search. Then, it will be selected one which is the best kernels that fits to the dataset from the three different kernels model. After that, this kernel is used to build the Model 1 with 13 features, Model 2 with 20 features, and Model 3 with 8 features and comparing the performance of three of them.

Model Optimization

The models that have been formed in the previous modeling stages are measured with accuracy by using 5-fold cross validation. The dataset is divided into 5 iterations by making randomly partitioned data into training and testing sets to guarantee that the present results are valid and can be generalized for making predictions regarding new data. In this research, this 5-fold cross validation is done in order to see how much the improvement of accuracy from the baseline model after the optimization of tuning the parameter by using Grid Search.

Table 4: Comparison between Baseline and Hypertuning Parameter Model of SVM

Kernel Model	Accuracy of 5-fold Cross Validation	
	Baseline	Hypertuning Parameter
SVM Linear	70%	72.85%
SVM RBF	72.85%	72.85%
SVM Polynomial	64.28%	70.00%

As it can be seen from the result of 5-folds cross validation, for SVM with Linear kernels, it shows an increase of accuracy as much as 2.85% after the parameter of the model is tuned with Grid Search. For, SVM with RBF kernels both for baseline and hyperparameter tuning model resulted in an accuracy of 72.85%. This shows no change since the optimization of the parameter by using Grid Search found that the parameter in the baseline has been optimal. While for SVM Polynomials, there's an improvement as much as 5.72% from the baseline. In general, it can be concluded that the optimization of hypertuning of each SVM model by using

Volume:6 Issues: 32 [March, 2021] pp. 1 - 14]
International Journal of Accounting, Finance and Business (IJAFB)
eISSN: 0128-1844
Journal website: www.ijafb.com

Grid Search has helped to increase the accuracy even though the improvement is not that significant for each model.

SVM Kernels Selection

In selecting the model, the 3 existing models of optimized SVM kernels are tested with 30% testing dataset in order to evaluate the model that has been built by training dataset. The accuracy is again used as it describes the fraction of predictions the model got right.

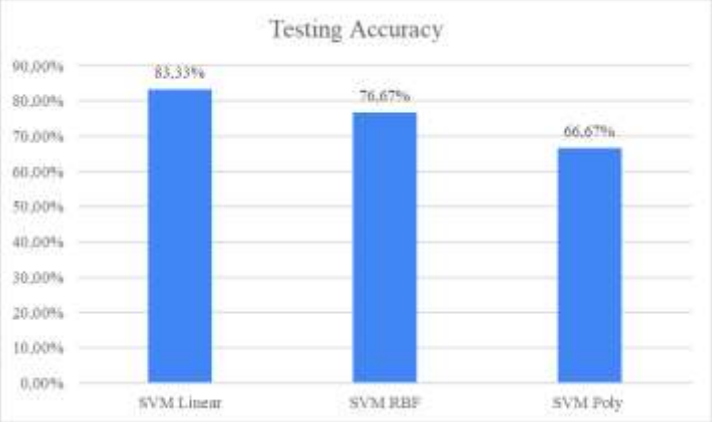

Figure 2: Accuracy of Different Kernel Model

As it can be seen from the result, SVM linear exceeds the accuracy of SVM RBF and SVM Poly. SVM Linear can hit the accuracy to 83,33% meaning that it has only 16.67% of error in predicting the incorrect classification. This accuracy is higher than the accuracy of SVM RBF and SVM Poly which is only 76.67% and 66.67% respectively. Yet, in order to select the model, it can't simply use the accuracy as the benchmark since it doesn't perform well with the imbalanced dataset where in this case, there is a difference between non-default class (68%) and default class (32%). Another benchmark will be used in selecting the best model. According to Zaki & Jr. (2014), ROC curve has invulnerability to class imbalance problems. This is caused by True Positive Rate (TPR) and False Positive Rate (FPR) which are not affected by the ratio of the amount positive class to negative class. Basically, the ROC curve formed will remain the same as a dataset that has a balanced number of classes or with datasets that have an unequal number of classes (Zaki & Jr., 2014). The performance could be examined on the area under the ROC that is called as Area Under Curve (AUC) which indicates the rate of successful classification by each model.

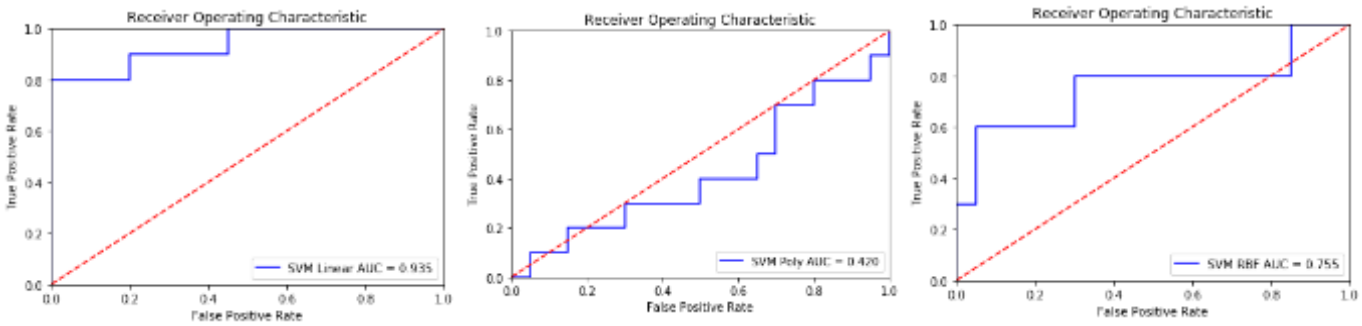

Figure 3: ROC Curve of 3 Different Kernels

Volume:6 Issues: 32 [March, 2021] pp. 1 - 14]
International Journal of Accounting, Finance and Business (IJAFB)
eISSN: 0128-1844
Journal website: www.ijafb.com

According to the ROC Curve, the SVM Linear has the highest AUC value which is 93.5% compared to the SVM RBF and SVM Poly which resulted in as much as 75.5% and 42% respectively. Hence, SVM Linear is selected as the best kernel to create the credit scoring prediction model since it performs better compared to the other two kernels.

Scenario Comparison

In order to see whether the adding of social media as additional features can increase the performance of credit scoring model, there must be comparison between the scenario with only historical payment and demographics features with the added one. Hence, there are 3 scenarios that will be compared which are:

Table 4: Scenario Comparison based on Features

Scenario	Features
Scenario 1	All Historical Payment and Demographics
Scenario 2	All Historical Payment, Demographics, and Social Media
Scenario 3	Selected Historical Payment, Demographics, and Social Media

All these three scenarios are all developed by using SVM Linear since the previous evaluation shows the result that the linear kernel is most fit to the dataset of this research. Next, to analyse the performance, the three scenarios are compared based on the area under the curve (AUC) of ROC and also the accuracy of the testing model.

Table 5: Comparison of Scenarios

Scenario	Number of Feature	Accuracy	AUC
Scenario 1	13 Features	73.33%	84%
Scenario 2	20 Features	83.33%	91%
Scenario 3	8 Features	83.33%	93.5%

According to the table, Scenario 2 and Scenario 3 has the same accuracy as much as 83.33% while Scenario 1 has only 73.33% from the data testing. On the other hand, Scenario 1 with 13 features comprised of all historical payment and demographics has 84% of AUC. After the addition of social media features in Scenario 2, there is an increment as much 7% of AUC becoming 91%. Meanwhile, it also can be seen that there is an increase of 2.5% of AUC in Scenario 3 becoming 93.5% after the feature selection is conducted by reducing the initial 20 features in Scenario 2 becoming only 8 features. From these results, it can be inferred that the addition of social media data to the model has significantly increase the performance of the model. The reason is mostly because the social media has the ability to describe the character of the borrowers. In the deeper understanding, social media data can explain how the personality trait of borrowers is in accordance with the real-life situation where this cannot be observed directly during the credit appliance process. Furthermore, social media can also indirectly tell about someone's social status and ties in the community. As a result, this social media can give signals to better understand the behavioural aspects of borrowers in which this can be used to develop the credit evaluation of borrowers where in this study these social media data is utilized to increase the performance of credit scoring model.

Conclusion

The application of data mining is done to recognize the characteristic patterns of customers who have default credit and non-default credit by building the model using Support Vector Machine (SVM). By using 100 borrower's data that is ranged from March 2017-January 2019 with the additional of social media features, it can be concluded that SVM Linear has the best performance compared to other kernels which are SVM RBF and SVM Poly, meaning that SVM Linear is the most suitable to user for the dataset of this research. SVM Linear shows the result of AUC as much as 93.5% while SVM RBF and SVM Poly are only 75.5% and 42% respectively. The addition of social media data as new features can increase the performance of the credit scoring model measured in AUC as much as 7% from the model which only includes the historic payment and demographic features. By this means, the social media data can improve the predictability of the credit scoring model.

References

Ala'raj, M., & Abbod, M. F. (2016). A New Hybrid Ensemble Credit Scoring Model Based on Classifiers Consensus System Approach. *Expert Systems with Applications*, 36-55.

Al-Mejibli, I. S., Abd, D. H., Alwan, J. K., & Rabash, A. J. (2018). Performance Evaluation of Kernels in Support Vector Machine. 2018 1st Annual International Conference on Information and Sciences (AiCIS), 96–101

Asian Bank Development Institute (2019). Fintech Development and Regulatory Frameworks in Indonesia. Retrieved May 25, 2020, from https://www.adb.org/sites/default/files/publication/532761/adbi-wp1014.pdf

Bank Indonesia. (2015). Profil Bisnis UMKM. Retrieved May 26, 2020, from https://www.bi.go.id/id/umkm/penelitian/nasional/kajian/Pages/Profil-Bisnis-UMKM.aspx

Bhatia, S., Sharma, P., Burman, R., Hazari, S., & Hande, R. (2017). Credit Scoring using Machine Learning Techniques. *International Journal of Computer Applications*, 161(11), 1–4.

Chalaris, M., Gritzalis, S., Maragoudakis, M., Sgouropoulou, C., & Tsolakidis, A. (2014). Improving Quality of Educational Processes Providing New Knowledge using Data Mining Techniques. *Elsevier*, 390-397

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. In C. Cortes, & V. Vapnik, Machine Learning (pp. 273-297). Boston: Kluwer Academic Publishers.

Cubric, M. (2020). Drivers, Barriers and Social Considerations for AI Adoption in Business and Management: a Tertiary Study. Technology in Society, 62, [101257]

Dellarocas, C., Wei, Y., Yildirim, P., & Van den Bulte, C. (2016). Credit Scoring with Social Network Data. *Marketing Science*, 35(2):234-258.

Fatemi, A., & Fooladi, I. (2016). Credit risk management: a survey of practices. *Managerial Finance*, 227-233.

Franata, R., Faturohman, T., & Rahadi, R. A. (2018). The Implementation of Credit Risk Scorecard Model to Improve the Assessment of Creditworthiness in A Peer-to-Peer Lending Company. *International Journal of Accounting, Finance, and Business (IJAFB)*, 3(13), 94-105.

Frohlich, H., & Chapelle, O. (2003). Feature selection for support vector machines by means of genetic algorithms. *Proceedings of the 15th IEEE international conference on tools with artificial intelligence*, pp. 142–148.

Greuning, H., & Iqbal, Z. (n.d.). Risk analysis for Islamic banks (English). Retrieved May 25, 2020, from http://documents.worldbank.org/curated/en/688471468143973824/Risk-analysis-for-Islamic-banks

Guo, G., Zhu, F., Chen, E., Liu, Q., Wu, L., Guan, C. (2016). From Footprint to Evidence: An Exploratory Study of Mining Social Data for Credit Scoring. *ACM Transactions on the Web (TWEB)*, b. 10, (4), 22:1-38.

H. Drucker, Donghui Wu and V. N. Vapnik. (1999). Support vector machines for spam categorization. *IEEE Transactions on Neural Networks*, vol. 10.

Hand, D., & Henley, W. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 523-541.

Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification. Retrieved July 1, 2020, from http://www.csie.ntu.edu.tw/ ~cjlin/papers/guide/guide.pdf

Li, Q., Chen, L. and Zeng, Y. (2018), The mechanism and effectiveness of credit scoring of P2P lending platform: Evidence from Renrendai.com. *China Finance Review International*, Vol. 8 No. 3, pp. 256-274

Khilfah, H., & Faturohman, T. (2020). SOCIAL MEDIA DATA TO DETERMINE LOAN DEFAULT PREDICTING METHOD IN AN ISLAMIC ONLINE PEER TO PEER LENDING. Journal of Islamic Monetary Economics and Finance, 6(2).

Kowalczyk, A. (2017). Support vector machines succinctly.

Malhotra, M., & Kanika, K. (2014). KDD-Knowledge Discovery in Databases. *International Journal of Innovatif Research in Technology*, 426-427.

OJK, D. o. (2019, March 01). *Kredit Macet Pinjaman Online Makin Tinggi, Apa Jadi Bom Waktu? (D. C. Syafina, Interviewer)*. Retrieved May 25, 2020, from https://tirto.id/kredit-macet-pinjaman-online-makin-tinggi-apa-jadi-bom-waktu-dhZM

Otoritas Jasa Keuangan. (2019). Laporan Kinerja 2019. Retrieved May 26, 2020, from https://www.ojk.go.id/id/data-dan-statistik/laporan-kinerja/Documents/Laporan%20Kinerja %20OJK%202019.pdf

Otoritas Jasa Keuangan. (2019). Perkembangan Fintech Lending (Pendanaan Gotong Royong) December 2019. Retrieved May 25, 2020, from https://www.ojk.go.id/id/kanal/iknb/data-dan-statistik/fintech/Documents/Perkembangan%20Fintech%20Lending%20Periode% 20Desember%202019.pdf

P2PMarketData. (2020). P2P Lending Explained: Business Models, Definitions & Statistics. Retrieved May 25, 2020, from https://p2pmarketdata.com/p2p-lending-explained/

Tan, & Phan. (2016). Social Media-Driven Credit Scoring: the Predictive Value of Social Structures. Retrieved June 7, 2020, from https://www.semanticscholar.org/paper/Social-Media-Driven-Credit-Scoring%3A-the-Predictive-TanPhan/2f1ce382e2be6ff6c70e2a43e0197d89426992c9

Trivedi, S. K. (2020). A study on credit scoring modeling with different feature selection and machine learning approaches. *Technology in Society*, 63, [101413]

World Bank. (2017). *Global Findex.* Retrieved May 25, 2019, from https://globalfindex.worldbank.org/sites/globalfindex/files/chapters/2017%20Findex%20full%20report_chapter2.pdf

Zaki, M. J., & Jr., W. M. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. New York: Cambridge University Press.