Loan Default Prediction Using Genetic Algorithm: A Study Within Peer-To-Peer Lending Communities

Lee Victor
MSc in Data Science and Business Analytics
Asia Pacific University of Technology & Innovation (APU)
Kuala Lumpur, Malaysia

Mafas Raheem
Academic, School of Computing
Asia Pacific University of Technology & Innovation (APU)
Kuala Lumpur, Malaysia

Abstract:- Peer-to-Peer (P2P) lending is a Fintech service that allows borrowers of any financial standing to be matched with lenders through online platforms without the intermediation of banks. Correct identification of probable defaulters is important for the longevity of the industry as the lender must bear financial risks should the borrower default, failure of which could result in loss of confidence and pulling out of the platform. However, with more information, it becomes difficult to determine the discriminatory features of the borrower. This study aims to develop a predictive model for loan default prediction in peer-to-peer lending communities. The predictive models were built using Logistic Regression, Random Forest, and Linear SVM with the selected feature set where Random Forest outperformed and achieved an accuracy of 92%. The significant fittest feature subset was obtained using a Genetic Algorithm and was evaluated using a Logistic Regression model. The Random Forest model could be used in the specified domain in this regard in future.

Keywords:- *Genetic Algorithm, Loan Prediction, Peer-To-Peer Lending, Predictive Modelling.*

I. INTRODUCTION

The consumer credit market represents one of the largest credit markets in the United States alone. The Federal Reserve Board of Governors has estimated an outstanding consumer credit of USD 4 trillion in Q1 2020 [1]. However, the credit market is not without its challenges. Lack of data and high default rates [2] has caused banks to turn away small-medium enterprises from seeking financial help. The less competitive nature of the banking system reduces the banks' incentive to provide more competitive rates to their customers thus maximizing returns in bank credit [3]. Furthermore, the lack of easy access to credit and stricter financial regulations may deter potential customers from presenting a greater barrier to entry to the credit market too [3].

The financial technology (Fintech) players have entered the credit market presumably for their ability to overcome these challenges [4]. Peer-to-peer (here on abbreviated as P2P) lending is a Fintech service that has gained prominence in recent years and known by other names like “debt crowdfunding” and “marketplace lending”, the platform initially began in the United Kingdom in 2005 [4]. The business model of P2P lending is the provision of loans similar to the service offered by brick-and-mortar banks. Borrowers (either individuals or businesses) apply to the P2P lending platform seeking loans [2][4]. Lenders (either individuals or a collective) decide if they should assume the loan along with its associated risks.

P2P lending platforms claim to offer a cheaper alternative to banks as its lack of intermediation, online presence and automatic screening of loan applicants means greater access to loans with reduced occurrence of information asymmetry [4][5]. The profitability of the platform should not be understated. To date, P2P loan platforms in the USA, Lending Club and Prosper captured 72.56% and 21.02% of the market share respectively [6]. The adoption of consumer loans on P2P lending platforms has been on an uptrend with a total of USD 48 billion in loans originated in just 12 years (2006-2018) [4]. The lucrative platform of P2P lending is poised to grow further with a projected growth to USD 150 billion by 2025 based on conservative estimates [7].

For all its perceived advantages, P2P lending is not without its series of challenges. The platform receives a portion of the loan volume originated in the form of revenue and as such is highly dependent on keeping a pool of active lenders [2]. As the lender bears the loan of the borrower, the borrower can default on the loan which may result in loss of lender confidence ultimately pulling out of the platform. This issue ties back to a core problem of information asymmetry where the borrower has better information on the lender’s ability to bear the loan [8]. P2P lending platforms attempt to mitigate this by providing historical information on borrower loans and their status in an attempt to improve transparency.

Besides, P2P lending platforms maintain a database of borrowers with different credit risks to level the information playing field between lenders and borrowers [9]. The rising prevalence of data analytics and algorithmic models has simplified the screening process with reduced cost.

The approval process of borrower credit involves evaluating numerous features some of which may not be relevant to the evaluation criteria and may lead to increased computation cost or decreased classification accuracy [10]. Researchers have in recent years developed and benchmarked the performance of many supervised machine learning models within the domain of P2P loan approval [11] [12] [13]. The development of probability default models has stagnated and that focus should turn to other modelling problems within the credit industry like data quality and feature selection [14] which inspired this study. Also, the feature selection methods have been used to improve the quality of results in clustering, regression and time series predictions [10]. A caveat in the concept of feature selection is the problem of large search space and the interaction between features. The relevance of a feature to the target variable could be enhanced or made redundant when paired with a complementing feature [10] [15]. Arbitrarily removing or selecting these features may neglect to find the optimal feature subset [10]. Thus, feature selection approaches often employ a search component that searches for the optimal feature subset(s) and an evaluation component that measures the quality of the feature subsets.

With business data growing larger, it becomes increasingly time-consuming and resource-intensive for financial platforms to quickly ascertain if the borrower is eligible for a loan. This study aims to develop a predictive model using a hybrid approach; that is, using a metaheuristic algorithm in the form of Genetic Algorithm (GA) for feature selection coupled with classification algorithms like Linear Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF) to build predictive models and choose the most suitable one.

II. LITERATURE REVIEW

Feature selection methods are generally categorized into filter and wrapper techniques [16]. The filter method evaluates each feature based on a statistical score like Chi-squared and information gain values where the highest valued feature used to get selected [16]. The filter methods are the best for large datasets as they are fast to implement [16]. Wrapper methods, on the other hand, evaluate a subset of features using a classification algorithm [16] [17]. Though wrapper methods are more accurate than filter methods, they are computationally expensive as the model needs to be called repeatedly and not suitable for large high-dimensional datasets [17].

Metaheuristic methods are also considered feature selection techniques that can search for the global optimal solution [18]. Recent literature has given prime focus to classification-based feature selection methods and feature selection blocks that utilize metaheuristic methods. The feature selection methods are categorized accordingly and discussed in separate sections.

A. Classification Based

The RF, Gradient Boosting Decision Tree and XGBoost were used as feature selection models and the features selected from those models were used in an LR model and their results were compared to other feature selection methods using F-score and Mutual Information [19]. Each model was evaluated using the area under the curve (AUC) and Kolmogorov Smirnov (KS) which measures the discriminative performance of the model. The models were run 10 times using 10-fold cross-validation and their results were averaged to get a robust prediction result. LR with RF and XGBoost as feature selection methods performed better than Mutual Information but worse than F-score. The author noted that the models provided an alternative to traditional feature selection methods and could be explored further as a feature selection method in P2P credit scoring.

A similar study was conducted using only RF models as a feature selection model where the model first evaluated the features of the dataset through importance ranking [20]. Subsequently, a correlation matrix was then applied to test the correlations of the selected features and 10 features were selected out of 12 in total. These features were then used to build decision tree models using CART (Classification and Regression Trees) and CHAID (Chi-square automatic interaction detector), Multi-layer perceptron and SVM with RBF (radial basis function) kernel and validated via 10-fold cross-validation.

The Restricted Boltzmann Machine (RBM) was used as a potential feature selection model which evaluated each feature using Root Mean Square Error (RMSE) and the features with the lowest values were selected [21]. Six classification models were built such as RF, SVM, k-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Linear Discriminant Analysis (LDA) and LR. LDA outperformed and chosen as the best model after applying 5-fold cross-validation on three separate credit datasets. The results proved its efficacy in the realm of credit risk analysis as the model can be used to speed up credit assessment.

Further, a statistical-based Minimum Redundancy Maximum Relevance (mRMR) and a wrapper based on least absolute shrinkage and selection operator (LASSO) as two potential forms of feature selection methods were applied on a dataset containing 33 variables and selected 27 [16]. With mRMR, the model was able to select the 10 most important features from the dataset. Both feature selection models were used as input to LR and RF classification models to determine its performance. LR models achieved similar performance to RF models but were easier to execute due to lower computational cost. The authors also observed that accuracy improved when the imbalanced dataset was under-sampled at the cost of higher false positives indicating that many good borrowers would be incorrectly classified as bad borrowers. As noted, under-sampling removes random observations from the majority class which may be important to the prediction process [22].

A two-stage approach was utilized in the feature selection process [23]. The authors first used recursive feature elimination (RFE) to select 30 features with the highest correlation to the response variable and then further reduced to 15 using a Pearson Correlation plot. The authors performed Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset. The selected features were used to build the RF, DT, LR and SVM models and RF was outperformed.

Ref. [24] developed a heterogeneous ensemble learning model comprising of three decision tree models such as Gradient Boosting Decision Tree (GBDT), XGBoost and LightGBM through a series of iterative training, cross-validation and heterogeneous ensemble learning. The model was also optimized via a change of model hyperparameters to improve the model's performance and through a feature selection method of which a learning model-based feature ranking method was used as it could execute the model learning phase and feature selection process in parallel. The decision tree based heterogeneous ensemble model outperformed individual classifiers on large datasets with a high rate of missing values. The authors concluded that hyperparameter optimization of XGBoost should be handled with scrutiny as the incorrect selection of hyperparameters could reduce model accuracy.

Studies in P2P lending recognized that the feature extraction methods could be either through statistical means or by carefully capturing the complex relationships in the data. However, this has the difficulty of scaling up to larger datasets. Ref. [25] explored the use of Convolutional Neural Networks (CNN) to predict repayment in P2P lending which is able to extract discriminative features and lending patterns in credit data. The study compared CNN's performance to other machine learning classifiers like KNN, SVM, MLP, DT and RF. CNN outperformed other classifiers in terms of accuracy and F1-score and maintained high performance after 5-fold cross-validation. CNN was also compared to other feature selection methods such as mutual information, information gain, chi-square statistics and RBM based extraction. The model performed similarly to RBM. The results reinforce the model being a good machine learning algorithm and feature selector.

B. Metaheuristic Based

Ref. [26] explored an extension of particle swarm optimization binary particle swarm optimization (BPSO) to perform feature selection and cross-validated with SVM for evaluation. To reduce the issue of early convergence and improve the fitness value of solutions, mutation operator, reset the best swarm and local search was used. The selected features were then used as input into tree-based extremely randomized trees (ERT) and random forest (RF) models. Classifiers with the BPSOVSM block achieved higher performance compared to classifiers without it particularly in terms of accuracy, AUC and execution time. The proposed feature selection block also had high performance given a smaller number of observations showing that it can achieve similar performance with a smaller subset of features.

Ref. [27] sought to develop a feature selection model using a competitive swam optimizer (CSO) a variant of particle swarm optimizer for its ability to handle large dimensional feature sets. The proposed model was combined with KNN and benchmarked against 6 separate datasets using 10-fold cross-validation to reduce the risk of overfitting. The proposed model achieved a lower average error rate and was able to select fewer features with the best fitness values at a faster rate outperforming conventional PCA-KNN method and PSO-KNN based variants.

Ref. [28] explored the use of feature selection using a binary competitive swarm optimization (BCSO) for feature selection. BCSO was compared with other metaheuristic models such as binary particle swarm optimization (BPSO), GA, Binary Differential Evolution (BDE) and Binary Salp Swarm Algorithm (BSSA). Performance was compared across 15 datasets where BCSO outperformed other metaheuristic models in terms of accuracy and ability to find the fewest significant features. Execution time is also very fast which lends itself well to real-world applications.

Similarly, Ref. [29] developed a predictive model combining Deep Learning Artificial Neural Network and a GA block to extract rules and plays the role of a filter. Observations that meet the rules of the filter will be classified immediately or they would be classified by the Neural Network block. The proposed model was benchmarked against other classification models across two credit datasets for a comprehensive test. To overcome the issue of overfitting, the model was run 20 times after which the type I and type II error rates were averaged. The proposed model outperformed other classifiers and was also noted to have low type I error rates across both credit datasets which can be useful in mitigating the risk of granting credit to bad borrowers.

Ref. [18] developed a novel loan evaluation model using RF with GA to maximize lender profit in the form of a profit score (RFoGAPS). The profit score considers both actual and potential returns and losses into the model evaluation criteria and was noted to be a better measure of performance on the loan evaluation model. The dataset was initially trained on the RF model and optimized accordingly. The GA was used to optimize RF with the objective of profit score maximization. The results of the proposed model using profit score as a metric using 10-fold cross-validation achieved a higher average profit score compared to other classification models.

Ref. [17] combined GA with filter techniques in the form of a hybrid genetic algorithm (HGA). The premise of the feature selection block was to select the initial subset of features using filter techniques like Gain Ratio, Information Gain, Gini Index and Correlation and then selecting the final feature subset using GA. The best-selected features were used in an ANN model. The proposed model was trained and tested on two separate credit datasets and its performance was compared to the regular GA-NN model using 10-fold cross-validation to obtain a robust result. The feature selection block managed to reduce the number of features while performing better in terms of accuracy over regular GA-NN models on both credit datasets. The authors concluded that the filter technique combined with GA to select the fittest features improved the classifier performance and disclosed it as a potential technique.

Ref. [30] approached the problem of feature selection in high dimensional datasets by using an Adaptive Potential Particle Swarm Optimization (APPSO) combining with a filter method Relief algorithm with a variant of PSO called Potential Particle Swarm Optimization (PPSO). Much like the work of Ref. [17], APPSO applied feature filtering and information entropy gain before PSO to remove more irrelevant features to improve the selection of features. The features selected by APPSO was used as input to KNN and tested on 10 high dimensional gene expression datasets. The proposed model was able to select 50% fewer features compared to PPSO and achieved higher performance in terms of accuracy.

III. METHODOLOGY

Figure 1 details the workflow of the proposed predictive model for the loan default prediction.

A. Data Collection

The dataset was from the P2P lending company named Lending Club available on Kaggle [31]. No personal information related to the borrowers were published in the dataset. The dataset comprises 2,260,701 observations and 151 variables and spans data from the years 2007 to 2018. The data from the year starting 2015 to the year ending 2016 was used as the dataset for this study.

B. Data Preparation

Data integration, data cleaning, data transformation and data reduction were carried out as part of the data pre-processing. The purpose of this stage was to improve the quality of the dataset thereby improving the accuracy and performance of the predictive model.

The dataset was analyzed to gain an understanding of the various features that are pertinent to the model development. Features such as occupation and borrower state might be used for data exploration to gain a better understanding of the demographics of the borrower base. The response variable loan status represents the current state of the loan at the creation of the dataset. This feature would be used to develop the predictive model using the other explanatory variables.

C. Exploratory Data Analysis

An Exploratory Data Analysis (EDA) was carried out to identify patterns, missing data and outliers in the dataset with the aid of visual plots and statistic measures. The purpose of performing EDA is to gain a keen understanding of the dataset and what features have the most influence on the response variable. A good understanding of the dataset enables the researcher to improve the predictive capability of the model. EDA used to be performed in conjunction with data pre-processing to ensure that any irregularities in the dataset are identified and treated accordingly.

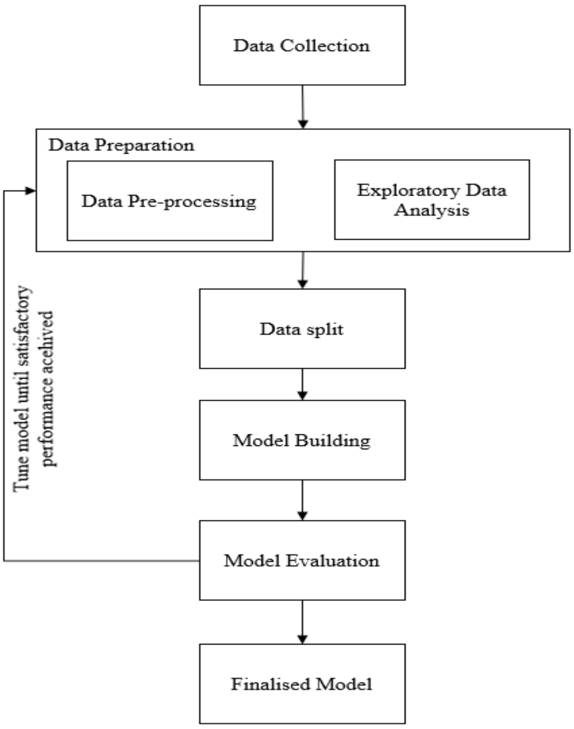

Fig 1. Workflow of proposed Loan Default Prediction Model

D. Model Development

The main aspect of the study was to explore the use of metaheuristic algorithms for feature selection. In particular, this study details a feature selection algorithm namely the Genetic Algorithm which selects the most important features in a dataset.

1) *Genetic Algorithm:* The Genetic Algorithm (GA) is a metaheuristic algorithm that was proposed as an evolutionary-based algorithm [32]. GA is based on Darwin’s principle of survival of the fittest where the fittest organism survives and goes on to produce even fitter offspring. The first step to using GA is to represent the individual solutions to a problem as analogous to a chromosome. This is done by representing the solutions as strings that can be randomly generated as an initial population [33].

The second step is to define a fitness function that will evaluate the fitness of the string. Higher values correspond to a fitter string. Conversely, lower fitness values mean the string performs less well on the problem. By initializing a starting population, each string is evaluated and the fittest strings are mated to produce a new generation that is fitter than the generation preceding it. However, it is also wise to consider weaker strings as it is possible that one extremely fit solution can be found in a few generations. This leads to premature convergence which is not desirable in GA [33] [34].

Ref. [33] details the following methods to select the parents to be mated:

- Tournament selection: Picking four strings from a population with replacement and selecting the fittest pair to be mated.
- Truncation selection: Picking a fraction of the best strings in the pool while ignoring the rest. A typical proportion would be to select and add 50% of the strings into the mating pool.
- Fitness proportional selection: Strings are selected through probabilistic means where the likelihood of a string being selected is proportional to its fitness.

Once the parent pairs have been selected, the process of breeding requires genetic operators. The most commonly used genetic operators are crossover and mutation operators and their functions are detailed as below:

- Crossover: A new string is generated comprising half of the genes from the first parent and half of the genes from the second parent. This is done by picking a random point in parent 1 and using the string of parent 1 up until the crossover point where parent 2 is used. This process generates two offspring where one string has the first part of parent 1 while the other string has the first part of parent 2. Crossover performs a global exploration in that the offspring created is different from the parent with the idea being that the offspring takes on the parent's best features.
- Mutation: As the name implies, a mutation involves a random change in the chromosome to yield a different result. In regards to this algorithm, the mutation is applied by changing the value of a string through some low probability, p. This technique promotes diversity in the population which can help in avoiding the local optimum solution of a population [35] i.e., the solution that is optimal to other similar solutions but is worse than the best possible solution of the problem (global optimum) [36] [37].

In summary, the steps to initiate GA are as follows [32]:

- Define the problem to be solved.
- Define a population of N individuals required for evolution.
- Define a fitness function with which to evaluate the individuals on.
- Perform crossover and mutation operators to generate offspring.
- Evaluate the fitness of the offspring.
- Select the best offspring based on fitness values.
- Stop if the criterion is reached. Otherwise, go to step 4.

2) *Logistic Regression:* Logistic regression (LR) is a variation of regression technique that predicts and explains a binary dependent categorical variable rather than a metric dependent variable [38]. In particular, LR explains the relationship between a nonmetric dependent variable and a set of metric or nonmetric independent variables.

The relationship between the dependent and the independent variables can be presented by the logistic curve as shown in Fig. 2 [38]. When the independent variable is at a low level, the dependent variables approach a probability of 0 but never reaches it. Likewise, at higher levels of the independent variable, the dependent variable approaches a probability of 1.0 but never reaches it. In this way, LR can identify how likely an object can fall within a particular group based on the probabilities of the dependent variable. This is particularly helpful in applications where the outcome is binary e.g. Yes/No. The LR models were recognised as the most appropriate models in deciding to grant credit to individuals and regarded as the industry standard in credit scoring model development [14] [38].

The probabilities of an object falling into either event can be rewritten as the odds ratio, that is, the ratio of the probability of two events occurring prob1 ÷ (1 – prob0). Taking into account the relationship of variables in the logistic curve, the LR model can be expressed by the following equations stated in (1) or (2):

$$Logit_i = \ln\left(\frac{prob_{event}}{1 - prob_{event}}\right) = b_0 + b_1X_1 + \cdots + b_nX_n \qquad (1)$$

Or

$$Odds_i = \left(\frac{prob_{event}}{1 - prob_{event}}\right) = e^{b_0 + b_1X_1 + \cdots + b_nX_n} \qquad (2)$$

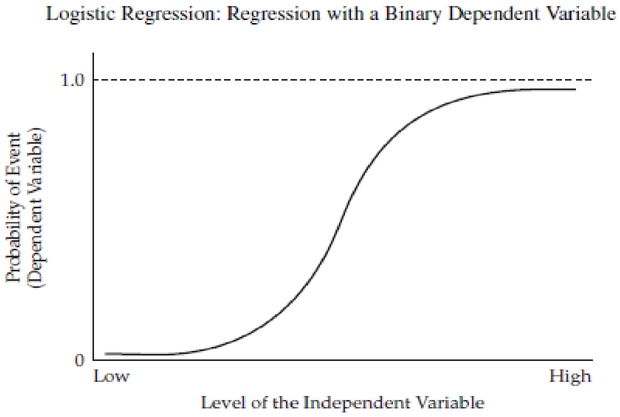

Fig 2. Logistic curve of the relationship between dependent and independent variables (Adapted from [38]).

3) *Random Forest:* RF is a supervised machine learning algorithm made up of an ensemble of decision trees. A decision tree model first starts at the root node and evaluates a range of attributes to make a decision. The nodes then branch off splitting the data into other decision nodes that indicate a potential outcome of the decision. This process continues until the model reaches a decision on every outcome and terminates (this is called the terminal node). Similarly, RF models contain a collection of decision trees that work in tandem to provide a prediction. RF models combine the principles of bagging and feature selection whereby each decision tree contains a randomly sampled number of observations and variables (with replacement) which are used to build the nodes. This helps to promote diversity in the decision tree models as

each decision tree is being trained on a different subset of data. The results of the ensemble of decision trees are combined averaged to improve the predictive accuracy of the dataset. RF has gained popularity as a predictive model due to its efficiency in handling high dimensional datasets and dealing with noisy and missing data.

4) *Support Vector Machine:* Support Vector Machine (SVM) is a supervised machine learning algorithm primarily used in classification and regression problems. The idea of SVM is to create a decision boundary, called a hyperplane with which the classes can be partitioned into. This boundary is established by noting the data points or support vectors between the two classes that maximize the distance of the hyperplane margin. SVM can handle both linear and non-linear problems where the former can be separated using a single straight line and the latter being that the dataset cannot be classified using a single straight line. It can handle nonlinear relationships by mapping them to a higher dimension space using nonlinear kernels. This makes it much more robust to overfitting though can be slow to train if the dataset has many variables or observations [39].

IV. DATA ANALYSIS

A. *Data Description*

The original dataset contained borrower data from Lending Club spanning the years 2007-2018. For the purposed of this study, data from the year starting 2015 to the end of the year 2016 was selected. The dataset contained 855,502 observations and 151 features. Initial exploration of the dataset revealed many columns having missing values of more than 30%, thus removed. This reduced the number of features to 93. With these large number of features, there is a possibility that the predictive model would become too complex. As such, reference to recent literature was referred to limit the feature space further [18] [23] [40]. The resulting features are tabulated in Table 1.

B. *Data Management*

The feature *employment length* was converted to a numerical variable with an employment length '< 1 year' to be 0 and '> 10 years' to be 10. The *home ownership* feature contained the level 'ANY' was not found in the Lending Club data dictionary. This level was removed from the dataset as it only accounted for 112 of the total number of observations.

The feature *purpose* contained 14 levels including 'Wedding' and 'Educational' which represented a small number of the total observations. The levels were then brought into the level 'Other'. A new feature called *earliest credit from issue date* was created by subtracting the date from the earliest credit line and the issue date to determine how old the credit line was at the date the loan was issued. The new feature was expressed in terms of months.

As the focus of the study is to provide a prediction for probable defaulters, the levels 'Late (31-120 days)', 'In Grace Period' and 'Late (16-30 days)' in the feature loan status was removed. The level 'Default' was then renamed 'Charged Off' as both levels shared similar definitions. For the numerical features, the *debt to income ratio* should only consist of positive values yet observations were containing negative values which were removed from the dataset. A right censor similarly seen in [40] was applied to the feature's delinquency in the last 2 years, inquiries in the last 6 months and public derogatory records. The rest of the numerical features contained outliers and were significantly left-skewed in their distribution. The top 1% of outliers was removed from each of the numerical features which improved the skewness and kurtosis values.

TABLE 1. Lending Club Attributes

No.	Feature	Description
1	Annual Income	The self-reported annual income provided by the borrower during registration.
2	Earliest Credit Line	The month the borrower's earliest reported credit line was opened.
3	Employment Length	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.
4	Home Ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER.
5	Inquiries in the last 6 months	The number of inquiries in past 6 months (excluding auto and mortgage inquiries). *Treatment:* Right-censor ≥ 3, meaning values more than 3 were set to 3.
6	Loan Amount	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
7	Loan Purpose	A category provided by the borrower for the loan request.
8	Open Account	The number of open credit lines in the borrower's credit file.
9	Total Account	The total number of credit lines currently in the borrower's credit file.
10	Term	The number of payments on the loan. Values are in months and can be either 36 or 60.
11	Debt to Income Ratio (DTI)	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
12	Loan Status	Current status of the loan.
13	Delinquency in the last 2 years	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years. *Treatment:* Right-censor ≥ 2
14	Instalment	The monthly payment owed by the borrower if the loan originates.

15	Verification status	Indicates if income was verified by LC, not verified, or if the income source was verified.
16	Revolving balance	Total credit revolving balance.
17	Revolving utilization	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
18	Public Derogatory Records	Number of derogatory public records. *Treatment:* Right-censor ≥ 3
19	Total Payments	Payments received to date for total amount funded.
20	Total Received Interest	Interest received to date.
21	Issue Date	The month which the loan was funded

C. Data Preprocessing

Missing values in the dataset were handled using mode imputation for categorical variables and mean imputation for numerical variables. The categorical variables were encoded using one-hot encoding to get it in the numerical format.

D. Feature Selection Block

GA was used with LR as the fitness function to determine the best feature subset. The best feature subset selected by the GA feature selection model was used as input to the LR model. GA was configured to have a crossover probability of 0.5 and a mutation probability of 0.2 to develop a string of fit individuals. The selection of parents to be mated was determined using a tournament selection whereby a set number of strings from the population was selected with replacement and the fittest pair was chosen to be mated. In this case, a tournament size of 3 was selected meaning 3 individuals participated in each tournament. Each offspring, in addition to being chosen for the mutation, had a small probability of having its attributes flipped (p = 0.05) to promote diversity in the population. The GA was initialized with a population of 50 and was run for 100 generations for a total of 5000 iterations.

At the end of the 100th generation, the GA resulted in a possible subset of features that performed well on the classifier. Generating more than one subset of features by the GA block was also possible which known as Multiple Optimal Solutions. In such cases, all the features were selected and a subsequent chi-square test statistic and multicollinearity check were performed to determine the relevance of the features concerning the output variable.

E. Data Normalization

A feature with a large range would be given more weight compared to a feature with a smaller range in the analysis. The normalization of data using min-max normalization was carried out to scale the data to enable the model to converge faster. Normalization was carried out on the independent features of the dataset. SMOTE technique was applied to the dataset to balance the classes using oversampling strategy.

F. Data Partition

The dataset was split into the ratio of 80% as training data and 20% as test data. A random state was added so that the data splitting is consistent and repeatable.

G. Model Evaluation and Assessment

The models were assessed using performance metrics such as precision, recall, specificity (Type I error), F1-score, accuracy, area under the curve (AUC) and Matthews Correlation Coefficient. These metrics can be calculated from the following equations based on the confusion matrix [33] [41]:

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

$$Specificity = \frac{TN}{TN + FP} \quad (5)$$

$$F1 - score = \frac{2TP}{(2TP + FP + FN)} \quad (6)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (8)$$

The metric precision evaluates the ratio of true positive instances out of the total pool of positively classified instances [33]. The metric recall (or specificity) is the ratio of true positive instances out of those that were classified as positive. Specificity, on the other hand, measures the ratio of true negative instances out of those classified as negative. Sensitivity and specificity measures are inversely proportional to each other and are used to determine the proportion of actual positive and negative cases that were correctly predicted respectively. The metrics of precision, recall, specificity and sensitivity have a value of 0 to 1 where 1 indicates a model with good classification ability. F1-score is the weighted average of precision and recall.

Recall and specificity can be used to construct the Receiver Operating Characteristic (ROC) curve. The closer the plot of the ROC curve is to the left-hand corner of the graph indicates that the classifier has performed well [33] [42]. The area under the ROC curve is known simply as AUC and represents the model's ability to discriminate between positive and negative classes [29] [43]. The AUC has a value of 0 to 1 with 1 representing a model that predicts the cases perfectly. To determine the quality of the predictions the accuracy of the models and Matthews Correlation Coefficient was used. Accuracy, as highlighted in the literature, has been a popular metric used to gauge the performance of the model. This would be used in conjunction with Matthews Correlation Coefficient (MCC) as a more robust metric to describe the confusion matrix and is used to measure the quality of the prediction. The MCC has a range of values from -1 to 1 with a value of 0.6 and above indicating a good classification ability [11].

V. RESULTS AND DISCUSSION

The dataset was used as input to the GA model which outputs a list of the most significant features. However, it was observed that the algorithm took a significant amount of time to run owing to a large number of observations and features present in the dataset. GA was run for 100 generations with a starting population of 50. Although this amounted to 5000 iterations, the model output 3 optimal feature subsets. Further steps were taken to check for the independent variable relevance to the dependent variable using chi-squared test statistics and reducing multicollinearity.

Each model was built using GridSearchCV with 5-fold cross-validation to find the parameters which yielded an optimal result as described in Table 2. The results show that LR and Linear SVM produced similar performance. However, the RF classifier performed the best with an accuracy of 92% and an MCC score of 0.77 which showed good classification ability. Precision and recall scores of the RF was also higher compared to that of LR and Linear SVM indicating good generalization ability.

Fig. 3 depicts the ROC curves of all three classification models. As stated in the preceding section, a ROC curve that tends towards the left-hand corner of the graph identifies a greater proportion of observations correctly (higher recall). A higher AUC value indicates the model's performance across all possible classification thresholds. As proved by Fig. 3 and Table 2, the RF model had good classification ability and outperformed the other models to better predict probable defaulters.

TABLE 2. Evaluation metrics using GA selected features

Classifier	Accuracy (%)	AUC (%)	MCC	Precision	Recall	F1-score
Logistic Regression	86	87	0.65	0.61	0.88	0.72
Random Forest	92	89	0.77	0.79	0.84	0.82
Linear SVM	85	86	0.64	0.60	0.88	0.71

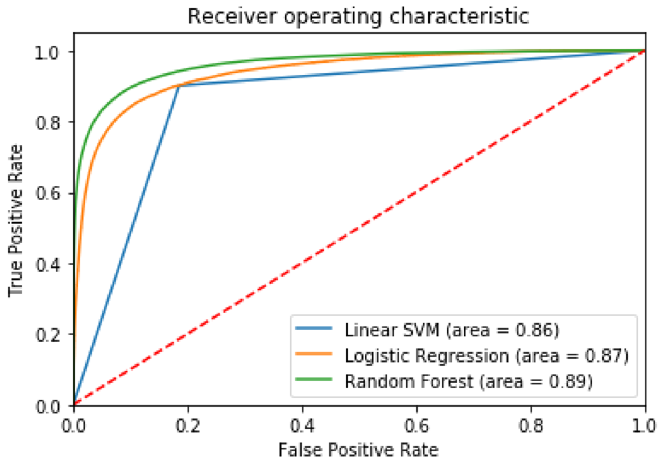

Fig 3. ROC curve of 3 classifiers.

Fig. 4-6 indicate the features that were useful in predicting the target variable. The features *total received interest* and *total payments* identified as more important by all three classifiers. Based on Fig. 4 and 5 respectively for LR and Linear SVM, a loan with a high total received interest would more likely default. The findings were intuitive as the higher total received interest on a loan indicates that the borrower is paying more interest to service due to low FICO scores and loan grade.

In contrast, *total payment* had a negative coefficient suggesting that a loan with high total payments would have a lower likelihood to default. Borrowers who have a higher total payment are being timely in their payments hence they would be less likely to default. RF model (Fig. 6) listed *total payment, total received interest, revolving balance, number of open accounts* and *term* as the top 5 features that contributed to reducing the weighted impurity when training a tree. Though the level of interpretability as seen in the feature importance plots for LR and Linear SVM was absent in the Random Forest model, which could only plot the magnitudes of the feature importance without a direction.

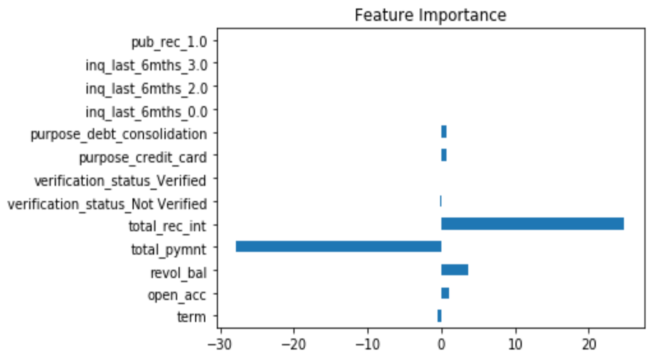

Fig 4. Feature Importance of Logistic Regression model

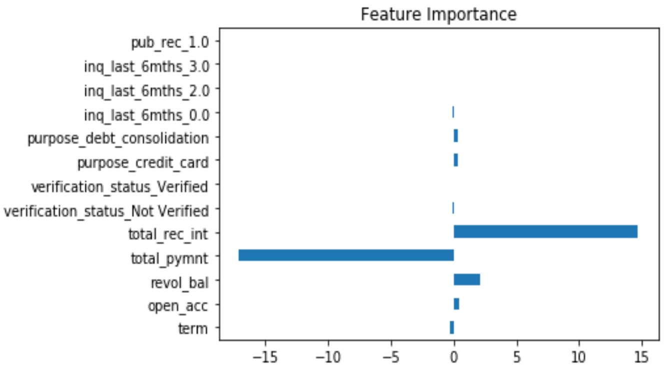

Fig 5. Feature Importance of Linear SVM model

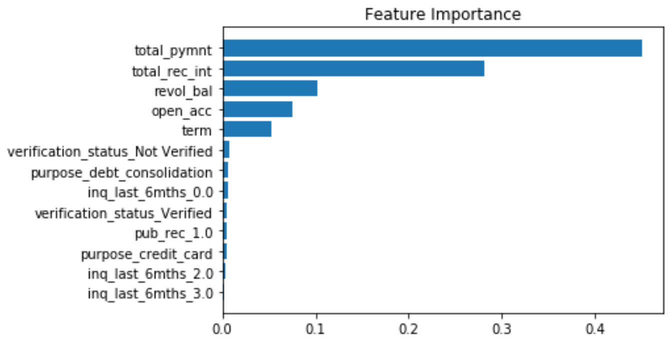

Fig 6. Feature Importance of Random Forest model

The performance of the models was compared to other related works in recent literature (Table 3). It was found that this study produced higher predictive performance in terms of accuracy compared to the other studies in predicting individuals who *Charged Off*. The findings suggest that further refinement could potentially help Fintech players as a supplement to existing systems to reduce investor loss thereby improving on trust and profitability of the platform.

TABLE 3. Comparison of results of past studies

Study	**Feature Selection**	**Model**	**Accuracy (%)**
Nguyen et al. (2019) [21]	Restricted Boltzmann Machine	LDA	**81.20**
		Logistic Regression	81.05
		ANN	66.08
		k-NN	72.55
		Linear SVM	76.60
		Random Forest	67.72
Setiawan, Suharjito and Diana (2019) [26]	Binary Particle Swarm Optimisation – Support Vector Machine	Extremely Randomized Trees	**64**
This study	Genetic Algorithm – Logistic Regression	Logistic Regression	86
		Random Forest	**92**
		SVM (Linear SVM)	85

VI. LIMITATIONS

Several challenges were faced during this study. GA was run for 100 generations with a starting population of 50. Although this amounted to 5000 iterations, the model resulted in 3 optimal feature subsets, a phenomenon of Multiple Optimal Solutions. This could be due to the insufficient iterations for the model to converge to an optimal feature subset. The feature selection block took nearly 4 hours to complete the process even though many iterations and a large number of observations and features present in the dataset.

Despite this, the feature selection block managed to reduce the feature space. It also presents an opportunity to select the best feature subset that would achieve the company's business objectives even though this would require a keen understanding of the business processes and intimate domain knowledge to make that decision.

Significant effort was made to optimize the predictive result through the GridSearchCV algorithm. Though the number of parameters explored was considerable but not comprehensive enough and could be further explored. The predictive models were limited given the time constraints and duration taken for each model to run. It would be prudent to build more predictive models to provide a more representative comparison of model performance on GA selected feature subsets.

VII. RECOMMENDATIONS

Future works into improving the feature selection and classification models can be summarized as follows:

A. Feature Selection Block

Tuning the model by modifying the mutation and crossover probabilities and altering the population size and number of generations to improve the convergence to an optimal solution. Alternatively, other metaheuristic algorithms like Differential Evolution and Artificial Bee Colony a variant of Particle Swarm Optimization can also be considered.

B. Classification Model

The models presented in this study is by no means exhaustive and presents an area of focus in future work. The models can be further optimized by modifying the model hyperparameters or by exploring other classification models like Naïve Bayes Artificial Neural Networks to further evaluate the classification performance on the selected feature set. Grid search methods are time consuming in that they evaluate a range of parameters one by one even if the particular parameter combination does not yield the best result. Automated hyperparameter tuning like Hyperopt could be explored as a potential alternative to parameter optimization techniques. These algorithms search for the best parameters through an informed approached whereby the model moves through the parameter space influenced results of the previous trials [44] [45].

REFERENCES

[1]. The Fed, (2020) *Consumer Credit - G.19*. [Online]. Available from: https://www.federalreserve.gov/releases/g19/current/. [Accessed: 26th July 2020].

[2]. Nemoto, N., Storey, D.J. and Huang, B., (2019) Optimal Regulation of P2P Lending for Small and Medium-Sized Enterprises. ADBI Working Paper 912. Available at SSRN: https://ssrn.com/abstract=3313999

[3]. Claessens, S., Frost, J., Turner, G. and Zhu, F. (2018) Fintech credit markets around the world: size, drivers and policy issues. *BIS Quarterly Review September 2018*. [Online]. pp. 29-49. Available from: https://www.bis.org/publ/qtrpdf/r_qt1809e.htm. [Accessed: 26th July 2020].

[4]. Balyuk, T., Financial Innovation and Borrowers: Evidence from Peer-to-Peer Lending (May 6, 2019). Rotman School of Management Working Paper No. 2802220, Available at SSRN: https://ssrn.com/abstract=2802220 or http://dx.doi.org/10.2139/ssrn.2802220

[5]. Bavoso, V. (2019) The promise and perils of alternative market-based finance: the case of P2P lending in the UK. *Journal of Banking Regulation*. pp. 1-15. Doi: https://doi.org/10.1057/s41261-019-00118-9

[6]. P2P Market Data (2020) P2P Lending & Equity Funding Statistics in the US (USD). [Online]. Available from: https://p2pmarketdata.com/p2p-lending-funding-volume-usa/ [Accessed: 28th July 2020].

[7]. PricewaterhouseCoopers (2015) Peer pressure: How peer-to-peer lending platforms are transforming the consumer lending industry. PricewaterhouseCoopers February 2015. [Online]. pp. 1-15. Available from: https://www.pwc.lu/en/fintech/docs/pwc-fintech-peer-pressure.pdf [Accessed: 24th July 2020].

[8]. Serrano-Cinca, C., Gutiérrez-Nieto, B. and López-Palacios, L. (2015) Determinants of Default in P2PLending. PLoS ONE 10(10): e0139427. Doi: 10.1371/journal.pone.0139427

[9]. Havrylchyk, O. and Verdier, M. (2018) The financial intermediation role of the P2P lending platforms, *Comparative Economic Studies*. 60(1), pp.115-130. Doi: https://doi.org/10.1057/s41294-017-0045-1

[10]. Xue, B., Zhang, M., Browne, W. N. and Yao, X. (2015) A survey on evolutionary computation approaches to feature selection, *IEEE Transactions on Evolutionary Computation*, 20(4). pp. 606-626. Doi: 10.1109/TEVC.2015.2504420

[11]. Bao, W., Lianju, N., Yue, K. (2019) Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Application.* [Online]. 128, pp 301-315. Available from: https://www.sciencedirect.com/science/article/abs/pii/S0957417419301472 [Accessed: 15th September 2020].

[12]. Turiel, J. D. and Aste, T. (2019) P2P loan acceptance and default prediction with artificial intelligence. [Online], p. 1-11. Available from: https://arxiv.org/abs/1907.01800 [Accessed: 28th January 2020].

[13]. Semiu, A. and Rehman, A. G. (2019) A boosted decision tree model for predicting loan default in P2P lending communities. *International Journal of Engineering and Advanced Technology*, [Online] 9(1), p.1257-1261. Available from: https://www.ijeat.org/wp-content/uploads/papers/v9i1/A9626109119.pdf [Accessed: 25th January 2020].

[14]. Lessmann, S., Baesens, B., Seow H. and Thomas L.C. (2015) Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research, *European Journal of Operational Research.* [Online]. 247 (1). pp. 124–136. Available from: https://www-sciencedirect-com.ezproxy.apiit.edu.my/science/article/pii/S0377221715004208 [Accessed: 16th August 2020].

[15]. Gheyas, I.A., Smith, L.S. (2010) Feature subset selection in large dimensionality domains. *Pattern recognition*, 43(2010), pp. 5-13. Doi: https://doi.org/10.1016/j.patcog.2009.06.009.

[16]. Chen SF., Chakraborty G., Li LH. (2019) Feature Selection on Credit Risk Prediction for Peer-to-Peer Lending. In: Kojima K., Sakamoto M., Mineshima K., Satoh K. (eds) New Frontiers in Artificial Intelligence. JSAI-isAI 2018. Lecture Notes in Computer Science, vol 11717. Springer, Cham. https://doi.org/10.1007/978-3-030-31605-1_1

[17]. Oreski, S. and Oreski, G., 2014. Genetic algorithm-based heuristic for feature selection in credit risk assessment. *Expert systems with applications*, 41(4), pp.2052-2064. Doi: https://doi.org/10.1016/j.eswa.2013.09.004

[18]. Ye, X., Dong, L.A. and Ma, D., 2018. Loan evaluation in P2P lending based on random forest optimized by genetic algorithm with profit score. *Electronic Commerce Research and Applications*, 32, pp.23-36. Doi: https://doi.org/10.1016/j.elerap.2018.10.004

[19]. Guo, W. (2019) Credit scoring in peer-to-peer lending with macro variables and machine learning as feature selection methods. 25th Americas Conference on Information Systems, (AMCIS) 2019, Cancun, Mexico, August 15-17, 2019. pp. 1-10.

[20]. Jin, Y., and Zhu, Y. (2015) A Data-Driven Approach to Predict Default Risk of Loan for Online Peer-to-Peer (P2P) Lending. *2015 Fifth International Conference on Communication Systems and Network Technologies, Gwalior, 2015,* pp. 609-613. doi: 10.1109/CSNT.2015.25

[21]. Nguyen, T., Khuat, T., Ngo, T., Nguyen, H. N. and Tran, M. (2019) Improve Risk Prediction in Online Lending (P2P) Using Feature Selection and Deep Learning, *IJCSNS International Journal of Computer Science and Network Security*. [Online]. 19(11). pp. 216-222. Available from: http://ijcsns.org/ [Accessed: 14th August 2020].

[22]. Fernandez, A., Garcia, S., Galar, M., Prati, R.C., Krawczyk, B. and Herrera F. (2018) Learning from Imbalanced Data Sets. Switzerland, Springer.

[23]. Zhu, L., Qiu, D., Ergu, D., Ying, C. and Liu, K. (2019) A study on predicting loan default based on the random for algorithm. *Procedia Computer Science*, 162,

pp.503-513. Doi: https://doi.org/10.1016/j.procs.2019.12.017

[24]. Zhou, J., Li, W., Wang, J., Ding, S. and Xia, C. (2019) Default prediction in P2P lending from high-dimensional data based on machine learning. *Physica A: Statistical Mechanics and its Applications*, 534 (122370). pp. 1-11. Doi: https://doi.org/10.1016/j.physa.2019.122370

[25]. Kim, J.Y. and Cho, S.B. (2019) Towards Repayment Prediction in Peer-to-Peer Social Lending Using Deep Learning. *Mathematics*. *7*(1041). pp. 1-17. Doi: 10.3390/math7111041.

[26]. Setiawan, N., Suharjito and Diana (2019). A Comparison of Prediction Methods for Credit Default on Peer to Peer Lending using Machine Learning. *Procedia Computer Science*, 157, pp.38-45. Doi: https://doi.org/10.1016/j.procs.2019.08.139

[27]. Gu, S., Cheng, R. & Jin, Y. Feature selection for high-dimensional classification using a competitive swarm optimizer. *Soft Comput* 22**,** 811–822 (2018). Doi: https://doi.org/10.1007/s00500-016-2385-6.

[28]. Too J., Abdullah A. R., Mohd Saad N. (2019) Binary Competitive Swarm Optimizer Approaches for Feature Selection. *Computation*. 2019; 7(2):31. https://doi.org/10.3390/computation7020031

[29]. Tran, K., Duong, T., and Ho, Q. Credit scoring model: A combination of genetic programming and deep learning, *2016 Future Technologies Conference (FTC)*, San Francisco, CA, 2016, pp. 145-149, doi: 10.1109/FTC.2016.7821603.

[30]. Huang, X., Chi, Y., and Zhou, Y. (2019) Feature Selection of High Dimensional Data by Adaptive Potential Particle Swarm Optimization, *2019 IEEE Congress on Evolutionary Computation (CEC)*, Wellington, New Zealand, 2019, pp. 1052-1059, doi: 10.1109/CEC.2019.8790366.

[31]. Kaggle (n.d.) *All Lending Club loan data.* [Online]. Available from: https://www.kaggle.com/wordsforthewise/lending-club [Accessed: 14th September 2020].

[32]. Holland, J. H. (1975) *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, Mich, USA, 1975.

[33]. Marsland, S. (2015) *Machine Learning an Algorithmic Perspective*. 2nd ed. Boca Raton, FL: Taylor & Francis Group).

[34]. Rocha, M. and Neves, J. (1999) Preventing Premature Convergence to Local Optima in Genetic Algorithms via Random Offspring Generation. In: Imam, I., Kodratoff, Y., El-Dessouki, A. and Ali, M. (eds) Multiple Approaches to Intelligent Systems. IEA/AIE 1999. Lecture Notes in Computer Science, vol 1611. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-48765-4_16

[35]. Shrestha, A. and Mahmood, A. (2015) Improving Genetic Algorithm with Fine-Tuned Crossover and Scaled Architecture, *Journal of Mathematics (2016). pp. 1-10.* Doi: http://dx.doi.org/10.1155/2016/4015845

[36]. Black, P. E. (2004a) local optimum, in *Dictionary of Algorithms and Data Structures*. [Online]. Available from: https://www.nist.gov/dads/HTML/localoptimum.html [Accessed: July 13th 2020].

[37]. Black, P. E. (2004b) global optimum, in *Dictionary of Algorithms and Data Structures*. [Online]. Available from: https://www.nist.gov/dads/HTML/globalOptimum.html [Accessed: July 13th 2020].

[38]. Hair Jr., J. F., Black, W. C., Babin, B. J. and Anderson, R. E. (2014). *Multivariate Data Analysis*. 7th ed. Essex: Pearson Education Limited.

[39]. Lantz, B. (2015) Machine Learning with R. 2nd ed. Livery Street, Birmingham: Packt Publishing Ltd.

[40]. Malekipirbazari, M. and Aksakalli, V. (2015) Risk assessment in social lending via random forests. *Expert Systems with Applications.* 42(10). pp.4621-4631. Doi: http://dx.doi.org/10.1016/j.eswa.2015.02.001.

[41]. Beyeler, M. (2017) Machine Learning for OpenCV: A practical introduction to the world of machine learning and image processing using OpenCV and Python. Birmingham: Packt Publishing

[42]. Zweig M. H., Campbell G. (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* 39(4). pp. 561-577.

[43]. Brownlee, J. (2016) *Machine Learning Mastery: with R Get Started, Build Accurate Models and Work Through Projects Step-by-Step*. [Online]. Australia. Available from: https://machinelearningmastery.com/ [Accessed: 14th August 2020].

[44]. Bergstra, J., Yamins, D. and Cox, D. (2013) Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. In *Proceedings of the 12th Python in science conference.* 13. pp. 13-19. Available from: https://conference.scipy.org/proceedings/scipy2013/pdfs/bergstra_hyperopt.pdf [Accessed: 8th February 2021].

[45]. KDnuggets (2019) *Automate Hyperparameter Tuning for Your Models.* Available from: https://www.kdnuggets.com/2019/09/automate-hyperparameter-tuning-models.html [Accessed: 8th February 2021].