MINIMIZING THE CREDIT RISK IN PEER-TO-PEER LENDING SYSTEM

ABSTRACT

Peer-to-peer lending (P2P) illustrates loan initiation method among investors and borrowers via virtual methods. This P2P structure permits investors to obtain higher rate of return on investment along with a possibility of the loan not being repaid in full. In peer-to-peer lending system have some advantages alongside with that it has some disadvantages as well such as to know correctly to assess the risk related with the borrowers. As peer-to-peer(P2P) Lending system are not so secured in nature of loan also comparatively with the original platform that makes a difficult problem in risk of the analysis. To reduce the this risk here I have used some traditional methods of which I find it useful and relative.

This project aims to analyze and reduce the threats linked associated to peer-to-peer (P2P) market of "Lending Club" Company. Machine learning and data transformation strategies were exercised towards investigating, analyzing and gaining fascinating associations between different attributes of the publicly accessible “Lending Club” 2013-2015 loan data. This helped in improving the accuracy of already existing prediction model that predicts defaulters subsequently helping potential investors determining the dynamics of the system by foreseeing the potential hazards associated with online P2P lending. Here I have used Machine learning as well as data mining methods that helps and enhanced to help to recognize the highly likelihood of borrower depending on some curtain parameters. Here I have evaluate the proposed come on a real life dataset of one of the largest and very famous Peer-To-Peer lending organization in United states of America. Here used traditional classification algorithm to illustrate the results, these results shows good performance of different borrowers, and their performance, loan status which can be using semantic data transformations. Here loan is assumed as ”Good ” only if it’s interest rate is expeditious and on the borrower as well ,whether borrower paid their loan on time or not. The algorithms are derived to enhance the favour of likely good loans borrower and also to recognize defaults or any risky credits.

INTRODUCTION

With the growth of P2P lending businesses, a large proportion of investors lean towards capitalize in non-commercial institutions. One can witness a sharp growth in the traditional purchaser borrowing loans through non-financial organizations. With the rise of P2P lending services, numerous investors convert themselves into personal banks. In peer-to-peer lending system ,small scale borrowers are as a separate and small organization and borrowers that are put down at the long tail of credits. These credits are mainly attracted to peer-to-peer lending due to unnecessary of the security for the loans and lack of the economical intermediaries.

Investors lend money keeping into account the reassurance of the loan being repaid by the borrowers. This implies that if the borrower is unable to pay in full, the amount lended by the investor, at that point the lender loses money. Investors has the rights to choose the amount of the capitals, also the interest cap and even the borrower.The peer-to-peer lending is not totally secure as it include the substaintial risk of defaulters to reduce this risk it is important to know and identify your borrower amog the list of unknown users. It is equally important to determine a good and a bad barrower but what is most important is to determine a good borrower ,also their potential and financing them first. As in Peer-to-peer income model is relative to the number of loans, size of credit, rate of interest etc but it sustain the losses on account of defaulters that will make total business unaccessible. Depending upon various important factor or on an attributes that are capable of diffrentiating between various loan application which will help to insure the minimum risk of credit. A good application has the heighest priority among all.

The impulsiveness of the borrower and the risk a lender takes to loan the applicant the money are two of the most critical queries presently existing in the lending industry. One can answer such question by regulating the rate of interest charged on the borrowed amount. The more uncertain the borrower will be, in terms of the amount borrowed and the time required to pay the debt, the higher will be the interest rate charged. Thus, we could then be able to distinguish the good creditors from the bad debtors taking into account the interest rates charged.

As here the number of attributes crdit are much smaller for every single record compare to number of credit records.Here i have used tree based classifier, this tree based classification which is a data mining technique will build a set of rules repetitively based on several input variables that are numeric categorical or outputs a class. Tree based classifier include tree like decision tree , Random forest ,Bagging < and ectra trees that helps and used to train prediction models for peer-to-peer Lending system.

“Lending Club” is a US based P2P lending organization that enables investors to lend money to needful borrowers through its business platform. “Lending Club” was stated as the biggest leading P2P lending system in December 2015. This company avoids the overheads and oddities of conventional banking system thus offering the borrowers enhanced interest rates and the investors a superior return on their investment.

Thanks to the farfetched assurances granted by the Lending Club, they were able to transform the traditional banking structure. As per the descriptions given by the borrowers, more than 80% of the loans borrowed at the Lending Club we used to repay the current loans or to pay their credit card debts where loan terms are either 36- or 60-month with static rate of interest and uniform payments. The borrowers are assessed and loans are then approved depending on credit scores, debt-to-income ratios (DTI), the length of the borrower’s credit history, and his/her current credit activity.

Lending Club has a process of allotting each loan a grade which ranges from “A” meaning the most significant with low rates to “G” meaning the least significant with high rates. For instance, as of 2018, an A-grade loan had an average rate of 7.62% while G-grade loans had an average rate of 30.89% [1].

In order to process the applications for loan through Lending Club there is a significant amount of data that needs to be collected to make accurate predictions. These include collection of variables corresponding to annual wages, job title, employment length, various financial trades in the last 24 months, bank balance etc. To find the optimal subset of these variables which are capable to accurately predict the potential defaulters is our principal necessity.

The project strives towards creating an evaluation model for P2P lending system based on an imbalanced classification algorithm for credit assessment. This project deals with improving the at now proposed “credit risk prediction” model in P2P lending through the utilization of complex classifiers, in particular to Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANN).

This project report arranged in such a manner that the section one give the brief introduction to Peer-to-Peer lending system to reduce the credit risk analysis. After in section two give the literature review, in which previous work shown. The section three is about the methodology and concept of different algorithm of machine learning, and later in section four and section five followed by conclusion and future work.

# LITERATURE REVIEW

Peer-to-Peer (P2P) lending system is presented as a modern electronic sensation in the business sector due to its ability for sustaining progressive efficient performance, which relied on a web-based framework (Berger & Gleisner, 2009; Wang, Greiner & Aronson, 2009) [2]. This system enables the desire and provision of capital over online services. Whether this innovation will end up unsettling the financial industry is a topic to discuss (Christensen CM, Overdorf M) [3], nonetheless the concept of P2P lending is unquestionably spreading briskly over the globe. “Lending Club” issued over $44.5 billion in loans in 2018, which is a 32.4% increase the previous year figure of $33.6 billion.

Applicants are either divided into two categories: borrowers and lenders in the P2P lending. One of the factors that distinguish Lending Club from its peers is that it allows the applicant to have an additional borrower if the primary borrower is uncertain to compensate the credit. Research shows that defaulters with less approval rankings had poorer achievement quotient with greater payments. Herzenstein et al. (2008) [4] discovered for analytical characteristics, viz nationality and gender of the applicant, had minute influence on access to loans. Yan Li (2014) [5] analyzed facts arising out of Paipai Mortgage, and his study demonstrated additional illustrative evidence relevant to debtors prompted to be credible of acquiring loans.

P2P loaning is a bilateral hub which is similar to the customary financing framework with its mainstream concerns (Klafft, 2008) [6]. Investors along with borrowers are an essential focus of all platform activities. As a result, a substantial amount of study has been concentrated on lenders along with the features responsible towards the progress of the lending system (Freedman & G.Z. Jin, 2008; Iyer et al., 2009) [7][8].

Peer-to-Peer lending system has easy and simple intermediaries with zero credit has low default risk, operation risk as well as policy risk, whereas more risk involved on the lenders according to the Li et al [9]. According to Wang and Greiner [10] recognize the fundamental issues of Peer-to-peer lenders in the credit risk of dropping on the investment, as loans are very unsafe with a no guarantee. Also credit risk is more in Peer-to-Peer lending system, as it leads to lenders investment is as high as in risk of defaults. To reduce this credit risk, Peer-to-Peer organization deploy different secured mechanism, like capital protection and recovery of areas [10].

According to the Lin et al. [11] has studied and presented peer -to-peer lending credit risk analysis of borrower in china. For this analysis he used logistic regression model to simplify the credit risk based on demographic features and their related loan information. As per his result, low risk of defaulters are of young age and mostly of them are female adults, who are professionally working, stable finance with low loan amount, highly educated and no default history.

As per Lee and Lee 12], Lenders in Peer-to-Peer lending are unprofessional investors and they need suitable skills to perform credit risk, in addition it has high credit risk of the lack of gage exposes. Due to needy of skills to assess credit risk and anxiety while choosing risky loans, many lenders manage to follow group of behaviour and accordingly finance loans with the high number of bidders.

Cinca et al [13], performed an experimental study on dataset lending club to examine credit risk in Peer-to-Peer lending. By using univariate and survival analysis Cinca et al studied first the parameters of predicting default risk. They found that few factors that are responsible for default risk which are loan grade, credit history, interest rate, income, loan purpose and borrowers type. To know the default risk they used logistic regression to the model and to get the loan grade for defaulters.

Cinca and Nieto has evolved a model of profit scoring for peer-to-peer lending ny using multivariate linear regression and by using decision trees. By comparing the results profit scoring with the credit scoring by using logistic regression method. After comparing both results divulged that the better performance get by using profit scoring models in recognizing profitable investments.

The research on debtors relies on subjective and quantifiable features to continually seek and enhance the defaulter’s indication system which will accurately predict the dependability of the borrowers.

Classification methods adopted in this model are "Artificial Neural Networks", "Support Vector Machines" and "K-Nearest Neighbors". The foremost objective to formulate these models has been to enhance “accuracy” rather than “precision” in order to identify good loans comparatively to the previous model which focused more on precision rather than the accuracy of the model.

As the investors struggle to identify the faulty prospects, the conventional loan evaluation techniques can be exploited in the P2P lending system. With the aim of recognizing deserving loan claims from faulty applicants, a wide range of data mining techniques are used in this project.

# EMPIRICAL ANALYSIS

The data utilized herein relates to loans issued by “Lending Club” from 2013 till 2015. “Lending Club” is responsible for all the information on “Issued” and “Non-issued” loans for helping lenders in choosing preferences.

The dataset deals with 656,724 loan records issued by "Lending Club", from 2013 till 2015 with a total of 115 attributes explaining the loan claims. The "Loan Status" characteristic that portrays the ongoing state of the application, has the associated values, "Issued", "Current", "Fully paid", "Default", "Charged off', "Late (16-30 days)", "Late (31-120 days)" and "In grace period". In order to eliminate these labels, they were transformed to a binary form, for instance, the claims such as "Charged off', "Default", "Late (31-120 days)" and "Late (16-30 days)" were regarded as "bad" or "defaulted" loans while "Current", "Fully Paid" and "In grace period" were tabulated as "good" loans and remaining labels were disregarded.[1]

Here grades are divided in group A to G according to the loan amount from 1000 dollars to 35.000 dollars respectively. Depending upon the grade and their loan amount interest rate are decide from 5.32% to 29% of range in ascending order. With highest grade G which has 31% and to the lowest it is 3% for grade A.

# METHODOLOGY

Data preprocessing and Cleaning

Data used to build the classification models was the exact same as the one used to build the previous model. The dataset was build using two files, for the financial year 2014 and 2015. The data consists of 235,629 and 421,095 records, 111 and 74 attributes respectively. The features which are mutual for both the datasets were kept in mind while framing the model. Since research done before the completion of the foundation model was completed in early 2016, the 2015 dataset had only 60,000 records which are grouped as "good" or "bad" credits, whereas the remaining were in "Issued" phase. 9.3% of 2014 dataset was additionally in an identical category and consequently disregarded. The preprocessed data contains of 10.9% credit defaults. The collective dataset has 279,169 with 70 attributes[1].

In order to clarify the distinctions about the improvements in performance over the previous model, only the attributes used in the previous model, namely “Loan amount, Interest rate, Installment amount, Annual income” were picked to build the new model while keeping the data values same.

The accompanying graphs show diverse relations and distributions of the attributes of data that is utilized to build the classification models. These graphs depict that the chosen variables exceptionally contribute to the ultimate conclusion made on the applicant’s decision.

Feature Selection

* Removing the attributes which are not useful or idle which contain the similar details which are need to be removed as well.’ Funded amount’ , ‘ amount funded by investors’ this attributes has the same data which is consider as extra and not needed are removed that is total 16 number of same attributes eliminated.
* Attributes that are missing values with highest missing value of 80% are not considered and also these attributes are found as optional as well. 10 to 15% of attributes which are left and has missing values are managed with the help of median of the available data which called as placeholder. Here in this model 25 such attributes are ignored.
* By removing unnecessary attributes which may result in inaccurate result need to be removed which are excessive, such as last payment date and last credit pull date also outstanding principle amount are only considered when borrower yet to repay the amount not before that. Due to such unnecessary attribute the pression and accuracy may increase {almost to 100%) and that will give undesirable result. Such 14 attributes get removed from the dataset for better future work.
* A natural method used on attributes which are left, for which it used the knowledge of important attributes and only keep the important one and removing rest the least important attributes. Such an 6 attributes which removed for classification and few of them are loan description, address state and public records.
* A tree based group of classifier uses decision tree considered as base learner which is also called Extra randomized forest classifier, which take samples through out the dataset while attributes are splitting at node. Here in feature set of every comination choose the sample at random which is ideally fit amongst all of them. It is important to spilt the attributes and it is despicable to train as node get split at randomly against to random forest algorithm.

**Classification**

* **Decision Tree**

In decision tree, decision tree classifier which uses supervised learning algorithm. This algorithm create a binary tree in which each node has an equal amount attribute split. By creating topdown decision tree which categorised data into sub set from their root node, this sub set contains similar set od data or values. Entropy is measured that is used to study the homogeneity, the entropy for the similar kind of sample set is 0 where as for non-similar sample value it is 1. By using Information gain which is calculated by entropy of parent node to the children’s weighted sum of entropy, here information gain value perform as the split between the each node.

**Table 1 Confusion Matrix of Decision Tree**

|  |  |  |
| --- | --- | --- |
|  | Predicted Positive | Predicted Negative |
| Positive  Conditions | 202586 | 46154 |
| Negative  conditions | 6051 | 24378 |

For Lending club dataset, above table give the attribute value of Confusion matrix for decision tree. Precision value comes near to 97% and the accuracy comes nearer to 82% for decision tree.

* **Random Forest**

Random forest classifier is supervised learning algorithm based on tree ensemble classifier. This classifier runs sample on the all of the decision tree that are generated and create more than one decision trees (which called as forest) for the random subsets of the data and also estimate the class with overhead frequency. Overfitting problem in decision tree can be avoided by using random forest. Random forest will help to increases the accuracy of model. In random forest value can may increase randomly for that it is necessary to split the data in subset. Similarly in decision tree here as well data is split in subset as ‘Interest rate’ and due to the randomness in subset rest of the attribute values may vary. Here the Precision value is near by 96% and accuracy value is at 88%.

**Table 2 Confusion Matrix For Random Forest**

|  |  |  |
| --- | --- | --- |
|  | Predicted Positive | Predicted Negative |
| Positive Conditions | 225545 | 23195 |
| Negative Conditions | 8909 | 21520 |

* **Bootstrap Aggregating or Bagging**

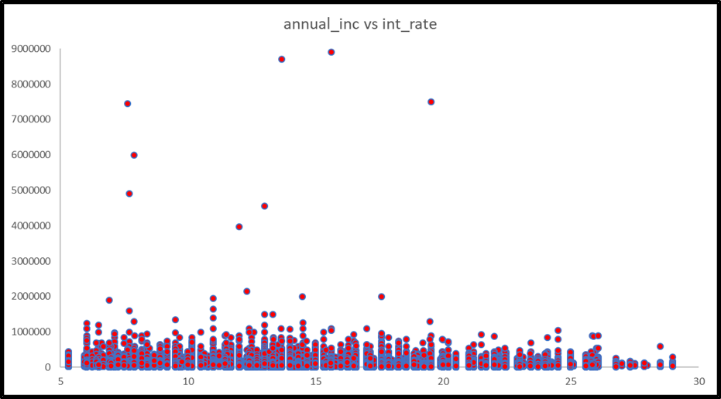
Bagging is a sampling method to get an approximate statistic rate of the sampling distribution by selecting a random subset value along with replacement, more than once and learning from it. It involves the bootstrapping of the primary datasets by using voting to get the better performance of another classifying algorithm. Overfitting problem can be solved in bootstrapping due its randomness in a nature.

**Table 3 Confusion Matrix For BootStapping or Bagging**

|  |  |  |
| --- | --- | --- |
|  | Predicted Positive | Predicted Negative |
| Positive Conditions | 224856 | 23884 |
| Negative Conditions | 8639 | 21790 |

Decision tree used here as the classifying algorithm for bagging. Here a precision received at nearly at 96.2 % and accuracy at 88.5%. Better performance of model can be achieved by random samples.

**Visualization on some attributes**

****

**Fig: Scatter plot showing “Annual income” vs “Interest Rate”**

Here the Scatter plot showing the plot of attribute annual income verses interest rate. Here for maximum rate of interest is for income which is more than or in between 700000$ to 900000$. For the maximum rate of interest is in the range of 5- 10% for income more than and equal to 1000000$.

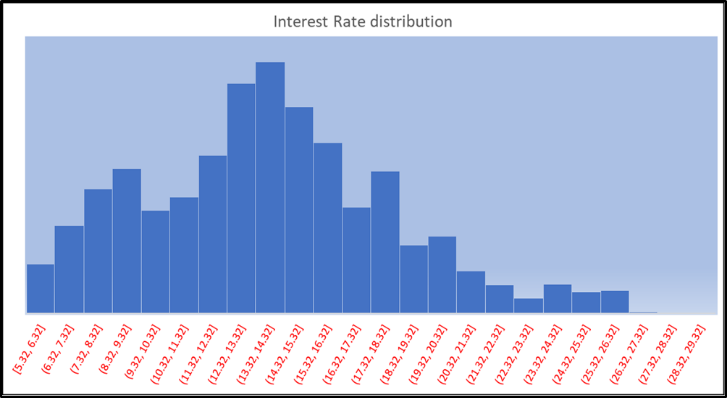


Fig: Histogram showing distribution of interest rates

Here the Histogram shows the distribution of interest rates for different rates. For maximum rate of interest it is showing at 13.32% to 14.32%. rate of interest depends upon the maximum income as well as borrower loan history.

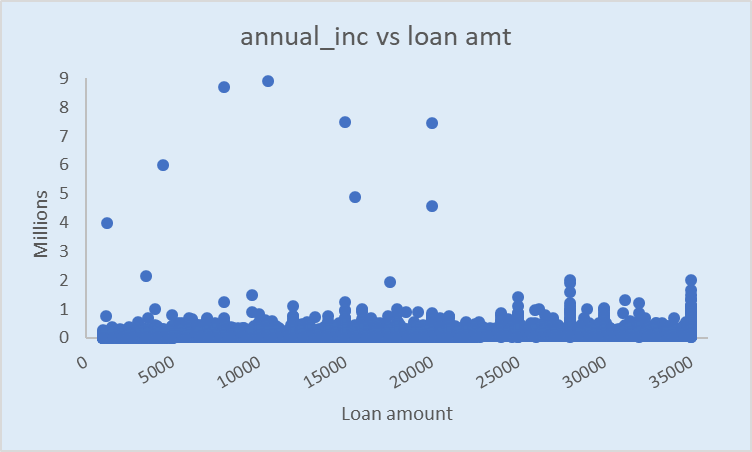
**

Fig: Scatter plot displaying Annual Income vs Loan Amount

Here Scatter plot showing the annual income verses loan amount, for maximum loan amount for borrower who’s annual income is in between 0 to 3 millions annually. The minimum loan amount taken by borrower who’s income is more than 6 millions.

# CLASSIFICATION METHODS

1. ARTIFICIAL NEURAL NETWORKS

Artificial Neural Network (ANN) is a simulated model that has a mechanism and approaches information in the manner of an actual living neural network in the human brain. ANNs are noticeable as biased coordinated diagrams in which artificial neurons are nodes and directed edges with weights are associations between neuron outputs and neuron inputs.

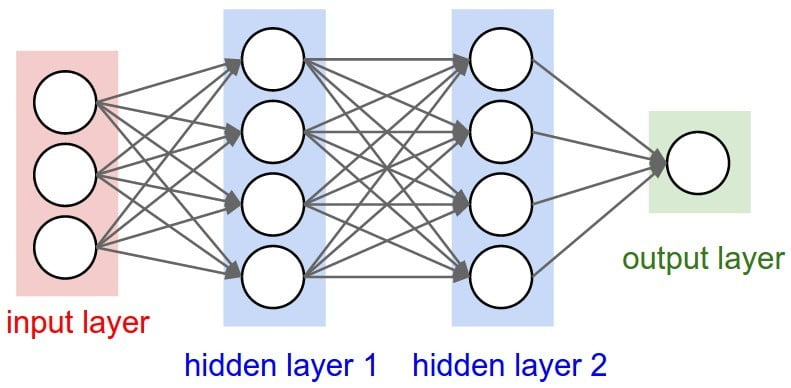
Artificial neural network is consist of input, output and nodes in between them. This nodes in between input and output in network called as Neurons. The connection between nodes is called as edges. At the input, it receives signal and processes and passes through neuron by edges between them and to output. In multilayer neural network , it consist of multiple layer and node between them and node at neuron. Each neuron receives signal process them and pass to other neuron in symmetric form through edges then to the output. This network work as in such way that output of one neuron can be input to the second neuron in next layer. Each node contains some weight which is added and process at the output. The sum of all nodes is available at output layer, the sum of all refer as propagation . Depending upon the weight of each node, the strength of the signal may increase and decrease. Signal passes through input and output layer in loop or for multiple times, neurons may have threshold only if the signal crosses the aggregated level. At different layer different transformation perform depending upon the input layer. In machine or in deep learning Artificial neural network mostly organised as multiple layer. The input layer receives external data and produce ultimate result at output and in between them may have zero or some hidden layers. In a one layer multiple neuron are present and those multiple neuron connected to single neuron present in next layer, this multiple neuron loop connection called as feedforward network in one after another neurons are connected.

Activation function plays important role in artificial neural network, depending upon the weight of every neuron and by calculating sum of all neuron activation function will decide weather to activate the neuron or not. Here the important factor is to introduce the non-linearity function. Activation function is important as it will simplify the network, reduce the complexity and and without activation function model can not learn the model and produce error. Also without activation function model will follow the simple linear function.

There are four types of activation function , one is Threshold Activation function used when binary input. This function activated by activating neuron when input is more or less than the threshold value then it is send to single neuron present in next layer. Second one is , sigmoid activation function which predict the probability of the output. The output value ranges between 0 and 1 only and output has curve shape of ‘S’, also called sigmoid shape function due to it’s characteristics. The sigmoid function may give the two sigmoid shape output as its output is differ. Hyperbolic tangent function is another Activation function that is more like similar to the sigmoid function but it has better performance than sigmoid. This function used when the value ranges between – 1 to 1 due to its non-linear nature and main advantage is that only for negative input it mapped negative output and similarly for zero input it mapped zero output.

The other activation function is Rectified linear units which is mostly used in Convolution Neural Network and Artificial Neural Network where the input ranges between 0 to infinity. For positive input it give positive output and otherwise it gives output as 0.

Neural networks, with their astounding ability to derive meaning from complicated or uncertain data, can be utilized to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. In ANNs there are multiple activations that can be applied to the hidden layers, in order convert a input signal to an output signal, for example, some of the frequently used activation function are sigmoidal or logistic function, tan-hyperbolic tangent function and relu-rectified linear unit function.



Artificial Neural Network model built for the Lending Club dataset has an input and output layer alongside with 2 hidden layers to obtain precise feedback. The activation function of tan hyperbolic sigmoidal function was favoured over sigmoid function not only due to the fact that the classification variable has only two classes but also because the range of the tanh function is between -1 to 1, which provides stronger gradient and a superior non-linearity over the sigmoid function.

When artificial neural network unable to decide the output on the previous knowledge it create the RNN that is recurrent neural network which help to decide the output. Independent decision are made by Artificial neural network depending upon the provided training set data and once it make the decision for use.

When the connection in between the layer are hidden and arbitrary it uses the CNN that is Convolution Neural Network. This network used in image processing and sound processing application.

The ANN model produced an accuracy of 90.075% and a precision of 96.60% with a group size of 500 and 150 epochs. The previously mentioned parameters were obtained by implementing hyperparameter tuning until the ideal accuracy was acquired.

*Table 1: Confusion matrix for Artificial Neural Networks*



1. SUPPORT VECTOR MACHINES

Support vector machines (SVM) are supervised learning models that are established on the concept of identifying a hyperplane that ideally differentiates the data into sections. Support vector machine algorithm can used in regression as well as classification model but it mostly prefer for classification objective. Support vector can perform non-linear classification called as karnel trick which divide the input in high dimension space. In support Vector machine algorithm it divide the same set of data into one space and another same set of data into other space by creating gap or line in between them. This create a hyperplane between two similar classes. There may be possibilities of more hyperplanes for similar set of data or classes. Dividing data set with better hyperplane which has largest distance from the data points of any class will result in less generalized error in accordance to classifier. In other words greater the hyperplane lesser is the error.

A better hyperplane is which divide the maximum set of classes, and if there is any nearest data points to the hyperplane or margin then that hyperplane is called as maximum margin hyperplane.

For labelled data into supervised algorithm is possible in SVM otherwise it will use Unsupervised algorithm.

SVMs effectively execute classification as well as regression examples and are overwhelming in higher dimension situations.

The aim of SVMs is to discover a hyperplane in a N-dimensional space that evidently characterizes the values. Support vectors are the components of the training set that would change the position of the separating hyperplane whenever removed and hyperplanes are decision boundaries that assist classify the values. Data points falling on either side of the hyperplane can be attributed to various classes.

A significant advantage of using SVM over ANN is that ANNs frequently overfits the model if training dataset goes on excessively long, implying that for any random pattern, an ANN may begin to consider the noise as a component of the pattern. Another advantage of using SVM is that they avoid generalization of the data which helps avoids the risk of over-fitting the model thus giving preferable outcomes than ANN.

*Table 2: Confusion matrix for Support Vector Machines*



The SVM model created for the Lending Club dataset was created utilizing “Linear” kernel, as the number of observations were severely huge and using a Gaussian kernel would have seriously influenced the speed of processing. This model produced an accuracy of 90.08% and precision of 95.69%.

3) K-NEAREST NEIGHBORS

K-Nearest Neighbors (KNN) algorithm is used in the grouping of unlabeled observations which works on the concept of assigning the unlabeled considerations to the most similar groups. Particularly, the model structure determined from the dataset. KNN is a distinct classifier that stores every single possibility and characterizes new situations based on comparability such as distance functions.

Another important factor in interpreting the effectiveness of the kNN algorithm is determining the parameter “k” which suggests the number of neighbors for the given algorithm . The value of k is generally taken as 2 if there are odd number of classes. The foremost step in KNN algorithm characterizes the estimation of the distance between a unique datapoint and the training data. There are various methods to compute this distance and for genuine input variables, the most prominent distance measure is Euclidean distance.

Euclidean distance: It is calculated as the square root of the sum of the squared differences between a new point and an existing point.

Other popular distance measures include:

Manhattan Distance: Calculate the distance between real vectors using the sum of their absolute difference.

Minkowski Distance: It is basically the simplification of Euclidean distance and Manhattan distance.

*Table 3: Confusion matrix for K-Nearest Neighbors*



The hyperparameters utilized in the algorithm were K (number of neighbors) = 37 and we used Minkowski distance metric with value of p as 2 i.e., KNN classifier used the Euclidean Distance Metric formula. The KNN model created for the Lending Club dataset produced an accuracy 90.071% and precision of 95.1%

# ALGORITHM PERFORMANCE ANALYSIS

Improvement in the accuracy of the prediction model using diverse machine-learning classification techniques such as ANN , SVM and kNN, when contrasted to Decision Trees, Random Forest and Bootstrapping techniques utilized in the previous model, was the primary objective of the project. Support Vector Machine displayed the best accuracy of 90.08% amongst all other techniques, which is 1.79% higher than the previous best classifier used in the existing model, Random Forest classification.

*Fig: Bar graph comparing the outputs of different classifiers*

*Fig: Precision and Accuracy graph*

Enhancing the accuracy of the prediction model, using Artificial Neural Networks, Support Vector Machines and K-Nearest Neighbors was the significant accomplishment of this project. The SVM model acquired the highest accuracy amongst every single other classifier’s though the Decision Tree classifier still stayed as the best classifier to get better precision.

# CONCLUSION

In this project, I have created prediction models utilizing machine learning techniques to foresee if a borrower will repay the loan based on historical data provided by Lending Club and to help investors to choose which investment strategy to pick.

The machine learning classifiers used in this project to obtain high precision and accuracy were Support Vector Machines, Artificial Neural Networks, and K-Nearest Neighbors. The accuracy acquired by Support Vector Machines classifier is 90.08%, which is 1.79% higher than Random forest (best classifier in previous model based on accuracy of prediction). This makes the Support Vector Machines predictor model better in recognizing the defaults. As the precision values of all the classifiers used in this project were below the previously best obtained value of 96.2%, it was eventually concluded that the Decision Tree classifier did a better job in obtaining optimum precision.

Numerous binary classification models generates the prediction of probability first and after that allocate the probabilities to 1 or 0 depending on the default threshold of 0.5. To improve the review of the model, we can utilize the probabilities predicted by the model and set threshold by ourselves. The threshold is set depending on several factors such as business objectives. Hence, the selection of cut-off value is one of the significant components provided that the investors are going to use this model to settle on which loan to invest or not, the decision of cut-off value will determine which applicants will get a loan and which not. In any case, it is good that investors can choose the percentage of bad rate that they are willing to acknowledge in their portfolio and depending on that they can choose very easily the percentage of the new loans that they want to finance.

After model building, testing, and data analysis, we can make conclusions that these models are ideal to locate the potential defaulters and, in this way, we can conclude that the credit risk in P2P lending system has been minimized.

# 

# FUTURE WORK

Although “Lending Club” is a prominent firm in the world-wide P2P lending business, yet it is still imperfect. A great deal of complications related to parameter selection while classifying the credits still demand additional enhancements.

Some information in the dataset have not been utilized in my model but is worth further examination. For instance, loan description and job title are composed as free text, which can conceivably provide sights into the inspiration and socioeconomic status of the loan applicant. Even though a variety of studies cross-examined the consequences of individual text characteristics (M. Greiner & Wang, 2007; Larrimore et al., 2009; Lin, 2009), minimal research has been done regarding the impact of the loan description on the evaluation of loan application. In my future work, I can extract relevant features from these texts using natural language processing (NLP).

Another extent of improvement can be obtained through the improvement in the Artificial Neural Network model. Combined with more data and multiple input nodes, a corresponding prediction with 3 output nodes (normal, attentive and warning) can be acquired which will help Lending Club in distinguishing the potential defaulters in a superior manner.

# REFERENCES

1. Kumar, S. Natarajan, S. Keerthana, K. Chinmayi, and N. Lakshmi, “Credit risk analysis in peer-to-peer lending system,” in Knowledge Engineering and Applications (ICKEA), IEEE International Conference on. IEEE, 2016, pp. 193–19

2. Berger SC, Gleisner F. Emergence of Financial Intermediaries in Electronic Markets: The Case of Online P2P Lending. Bus Res. 2009;2(1): 39–65.

3. Christensen CM, Overdorf M. Meeting the challenge of disruptive change. Harvard Bus Rev. 2000; 78(2): 66–77.

3. GAO Ruiqiong, FENG Junwen. An Overview Study on P2P Lending. Int Bus Manage. 2014;8(2): 14–18.

4. Herzenstein, M., Andrews, R.L., Dholakia, U., et al., 2008. The Democratization of Personal Consumer Loans? Determinants of Success in Online Peer-to-Peer Lending Communities. Boston University School of Management Research, Paper

5. Li, Yan, Gao, Yijun, Li, Zhenni, Cai, Zihao, Wang, Bingting, Yang, Yuxuan, 2014. The influence of borrower’s description on investors’ decision–analyze based on P2P online lending [J]. Econ. Res. J. (s1), 143–155.

6. Klafft, M. (2008). Peer to Peer Lending: Auctioning Microcredits over the Internet. Proceedings of the 2008 Int’l Conference on Information Systems, Technology and Management (pp. 1-8). Dubai: IMT. Retrieved from <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1352383>

7. Freedman, S., & Jin, Ginger Zhe. (2010). Learning by Doing with Asymmetric Information : Evidence from Prosper.com. College Park, MD. Retrieved from <http://kuafu.umd.edu/~ginger/research/Freedman-Jin-Feb2010.pdf>

8. Iyer, R., Khwaja, A. I., Luttmer, E. F. P., & Shue, K. (2009). Screening in New Credit Markets Can Individual Lenders Infer Borrower Creditworthiness in Peer-to-Peer Lending? Management. Cambridge, MA.

9. References for coding templates and concept intuition in python: https://www.udemy.com /machinelearning/learn/v4/t/lecture/5714404?start=0

<https://www.udemy.com/machinelearning/learn/v4/t/lecture/5714406?start=0>

<https://www.udemy.com/machinelearning/learn/v4/t/lecture/6760380?start=0>

<https://github.com/robertmartin8/udemyML>