Loan Default Prediction using Ensemble Learning

Rohan Maheshkumar Aswani
dept. of Computer Science
University of Windsor
Windsor, Canada
aswani@uwindsor.ca

Rakshana Bagavathi
dept. of Computer Science
University of Windsor
Windsor, Canada
bagavat@uwindsor.ca

Aneerban Chakraborty
dept. of Computer Science
University of Windsor
Windsor, Canada
chakrab6@uwindsor.ca

***Abstract*—Credit Analysis is a key aspect of any credit lending business. With an astronomical increase in consumer need and commercial competition, creditors are getting more conscious about risk analysis and management when it comes to predicting whether an applicant will default or not. The paper aims to juxtapose various machine learning models for loan default prediction on a highly imbalanced dataset. Support Vector Machine, Random Forest, Neural Network, Balanced Bagging Classifier, Stacked Ensemble Learning Model is used alongside re-sampling data to get the desired results. ROC AUC scores are used along with Confusion Matrix to validate and compare the models. Stacked Ensemble Learning model along with over-sampling gave the best results on the chosen dataset.**

Keywords—Binary Classification, Feature Selection, Ensemble Learning, Machine Learning, Default Prediction

INTRODUCTION

With the current scenario of global finance, Risk management and analysis has been an important issue [1–4]. Skyrocketing credit card debts are a problem for banks and card holders alike. Taking the example of the US, the country hit a record high of $930 billion in the last quarter of 2019, with younger Americans having the highest delinquency rate [5]. There is often a misconception with credit card debt, loans and the economy. It is true that consumer spending drives the economy. However, what is often overlooked by people is that this growth eventually stops when households operate at a loss. Therefore, it is crucial to keep a check on the credit card and loan defaults.

With the exponential growth of consumer-level data, it is becoming more profitable to use big data analysis for extracting high-level features from this humungous amounts of data available for statistical and Machine Learning analysis. The aim is to predict the probability of a person defaulting, even before they are lent any money. This can be done by a analysing the information of a person. Accurate estimation is beneficial for both creditors and applicants. The creditors benefit from reduced losses and applicants can avoid any overdue that they might be unable to pay. This kind of default prediction is done on all levels, from banks to retailers, with the key objective of reducing losses due to delinquency. Although delinquency rate is on a steady decline since 2012, it is still not insignificant, at roughly 2.3% in the first quarter of 2022 [6] in the US.

Before big data analysis and machine learning algorithms had taken over, traditional methods for accessing whether to grant a credit to a person of not used human judgement of risk, based on experience. However, the increasing consumer demand and cut-throat commercial competition does not permit these traditional methods to be relevant in the current scenario. There exist sophisticated statistical models for making these decisions. The commonly used "Credit Scores" are universal metrics to judge the creditability of person [7]. Credit Scoring algorithms have existed since 1950s, with popular ones from Fair Isaac Corporation (FICO) being widely used. Credit score is not just used for its original purpose, instead it also used by insurance companies for setting insurance rates and even by employers for pre-employment screening as character analysis. There are even online dating websites that claim to match users with high credit scores. These credit scores act as a psychological pressure on people, incentivising them to pay back their debts and penalizing them for not doing so. Even with world-wide implementation of credit scores, these have flaws and limitations. Apart from being heavily biased based on minorities and income levels, these are not always accurate for predicting defaults.

In this project, we propose an Ensemble learning model that uses several machine learning algorithm on "Loan Default Prediction - Imperial College London" dataset from Kaggle. We have used this dataset to perform Binary classification: If an applicant will default or not.

LITERATURE REVIEW

Credit Scores and Credit predictions are a key aspect of today's financial system. This has brought focus to extensive research is credit risk analysis [8-11]. The ever-increasing demand and reliability of financial institutions on loan lending calls for a constant improvement in models for default prediction and credit scores. Some existing models are as following.

Logistic Regression

Being fairly simple and fast, Logistic Regression is one of the most commonly used statistical analysis algorithm used for credit analysis. It has an easy implementation and gives in a sturdy performance. Loan default detection is a probability-based question and thus the expected output should be any value between 0 and 1 which is conveniently provided by Logistic Regression [12].

Logistic regression has been used previously by many researchers. One of the research topics being to improve the accuracy of the coefficients of the logistic regression algorithm [13]. Kemalbay, G. and Korkmazoğlu, Ö.B have proposed a two-step solution wherein first a categorical principal component analysis is used to address the problem of multi collinearity. Second, is using the uncorrelated features for prediction. Using this algorithm of utilizing the main components for explanatory features has been proven beneficial to correctly classify housing loan approval data with 91.1% accuracy.

Evaluation of scorecards is the most basic way of determining credit score. This system has many disadvantages and hence, Sohn, S.Y., Kim, D.H. and Yoon, J.H. [14] proposed a fuzzy logistic regression credit score model used

in predicting loan default. This takes into consideration the linguistic expressions which are basically the attributes that are associated with the technology in predicting loan default using fuzzy inputs. Credit Scoring is an example of cost-sensitive classification meaning that the costs of misclassification differ between cases. Bahnsen, A.C., Aouada, D. and Ottersten, B. [15] proposed a different example-dependent cost matrix which is an additional input to logistic regression algorithm.

In a similar study, researchers Sohn, S.Y. and Kim, H.S., [16] have used the random effects logistic regression model which accommodates the discrete attributes as well as the unreliable attributes that are not accommodated by these discrete attributes. These attributes are based on both financial and non-financial factors. This model of incorporating all the deterministic and non-deterministic plausible elements affecting the funding decision is proven to be beneficial for the Korean funding agency. Another case study is about the loan defaulter prediction system of Ghana in the microfinance sector specially keeping into consideration that about 30% of people living there are under the poverty line index. Agbemava, E., Nyarko, I.K., Adade, T.C. and Bediako, A.K. [17] have studied and identified the risk factors which varies according to individuals. They implemented logistic regression model with 86% accuracy.

Support Vector Mechine

Support Vector Machines (SVM) is based on statistical learning theory [18]. It has been widely used for classification problems because it follows either one-against-one or one-against-all for classifying objects into target labels [19]. SVMs provide good classification results even when trained on limited number of samples contained in the training data.

Loan default detection in calculated based on a collection of standards; however, as of today many factors affect the credit score of an individual. Researchers Moula, F.E., Guotai, C. and Abedin, M.Z [20] have taken into consideration all these factors in determining the Credit Default Prediction (CDP). They used SVM algorithm on selected criterions and also by encompassing some new factors that affect the performance of their model. They then compared their performance statistics with six different datasets. They concluded that performance of SVM model is robust and better with an accuracy of 85.32% as compared with other classifiers.

In contrast to the classic SVC wherein two parallel supporting hyperplanes are constructed in order to segregate the classes, researchers Shao, Y.H., Chen, W.J. and Deng, N. [21] have worked on constructing two nonparallel hyperplanes. This overcomes the disadvantage of lack of steadiness when neither the relative distance between hyperplanes not a datapoint become visible concurrently during training process. Therefore, they proposed a nonparallel hyperplane SVM (NHSVM) for binary classification with modelled consistency between the training and prediction procedures. The NHSVM model [21] solves a single quadratic programming problem which is persistent between training and prediction operations and fabricates two nonparallel hyperplanes. These hyperplanes are used to cluster the datapoints in accordance with the alike classes. The accuracy of the NHSVM model is 90.68%, surpassing the accuracy and performance of other algorithms.

Supporting the previous study of Korean funding scheme to SMEs [16], it has been brough to notice by researchers Kim, H.S. and Sohn, S.Y. [22] that a better model that detects the score accurately should be implemented in order to overcome the high default rate that had been reported previously. The proposed model is SVM which takes into consideration various essential features. The accuracy of their SVM model with cross-validation is 66.16% on their dataset which is performing better compared to logistic regression and back-propagation neural network. Another similar study by researchers Eweoya, I.O., Adebiyi, A.A., Azeta, A.A. and Amosu, O. [23] have proposed a machine leaning based approach using SVM. This model takes into account the hidden features in the data which in other case cannot be generally seen by past records. This model resulted in a 81.3% accuracy.

Hitherto, the performance of SVM has been better as compared to logistic regression and in some datasets it has even outperformed the back-propagation neural network model as well in detecting any kind of loan default prediction. Some research has even been conducted to combine SVM with other algorithms in order to optimize the overall prediction process. Vimala, S. and Sharmili, K.C. [24] have combined Naïve Bayes with SVM (NBSVM) which resulted in a much faster execution time than each of them when run individually. This model has inherited the robust and simplicity of Naïve Bayes algorithm and the precision and fast execution time of the SVM algorithm yielding 77% accuracy.

Decision Tree

Decision Tree is one of the most versatile and robust classification algorithms. Due to its versatility and robustness, it is one of the most widely used algorithms for ensemble learning. Decision tree outperforms other classification algorithms especially when the data is imbalanced. In decision tree, each leaf node is the class target and the internal nodes represents the conditions that helps in identifying to which class the datapoints belong to [9]. This tree structure of decision tree follows recursive portioning algorithm for classification [25].

Research in comparing decision trees performance compared to other classification algorithms has been done extensively in predicting the loan defaulters. A study on comparing logistic regression, neural networks and decision tree algorithm has been done by Zekic-Susac, M., Sarlija, N. and Bensic, M. [26] wherein they specifically used CART decision tree. The methodology of their approach is that it takes one input at a time in the function and constructs a binary-tree by breaking every node. Each sub-tree is constructed and one is selected in accordant with its error rate. Another research by Zekic-Susac, M., Sarlija, N. and Bensic, M. [27] have used decision tree for predicting credit score and achieved an accuracy of 74% which is better as compared to other classifiers.

Ensembling is an efficient way to improving the prediction performance of any prediction algorithm [28]. There are two ensemble methods widely used in predicting the credit score, namely, parallel ensemble and sequential ensemble. Researchers Xia, Y., Liu, C., Li, Y. and Liu, N. [29] have proposed a credit score predictions system based on sequential ensemble using XGBoost. This model has accuracy 84.65% which surpasses all the other models. Their result not only provides the accurate prediction, but also anticipates the useful

features and decision chart for improving the credit score system.

The functionality of the conventional bagging technique on other classifiers is that same features are used in training every model. Researchers Zhang, D., Zhou, X., Leung, S.C. and Zheng, J. [30] have proposed a vertical bagging decision tree model (VBDTM) that is performing better than the conventional bagging technique wherein the model obtains the collection of classifiers through predictive features as input and all the training sample datapoints as well as features take part in the learning processes of each classifier model, hence naming it a vertical bagging method. They tested their performance on two credit datasets, namely German credit and Australia credit, where the accuracy using the vertical bagging decision tree model is 81.64% and 91.67% respectively [30]. Furthermore, they used majority vote and weight vote as a two vote strategy to build their model and improve performance of prediction for credit data. They concluded that eventhough comparable, weight vote has performed slightly better on their dataset and hence it is evident that prediction accuracy is affected by the size of decision tree.

Random Forest

Due to the imbalanced nature of datasets for loan default predictions, random forest classifiers are favored for this application. Lin Zhu et al. [31] and Nazeeh Ghatasheh et al. [32] adopted random forest for loan default prediction. Zhu concluded that random forest classifier performed better than other models with accuracy of 98%, compared to just 73% of logistic regression, 95% of decision tree and 78% for SVM. Ghatasheh also concluded that random forest classifier is one of the best for credit risk prediction due to its competitive accuracy and simplicity.

K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a commonly used machine learning algorithm used for both classification and regression problems. It is simple and easy to understand yet gives highly competitive results.

Research is done extensively on applying KNN to classification problem, especially for predicting credit score. Great results have been achieved when using the KNN algorithm. Studies have also been done to optimize the performance of the KNN model. Researchers Mukid, M.A., Widiharih, T., Rusgiyono, A. and Prahutama, A. [33] have used weighted k nearest neighbour (WKNN) method by considering the use of some kernels. We use credit data from a private bank in Indonesia. The result shows that the Gaussian kernel and rectangular kernel have a better performance based on the value of percentage corrected classified whose value is 82.4% respectively.

Neurral Networks

The limitations of the classical machine learning models can be somewhat tackled using Neural Networks. Neural Networks have been a major focus of research in the recent decades across the globe. They have the capability of approximating complex function well. A neural network generally consists of three kinds of layers, Input layer, hidden layers and output layer. As the name suggests, the input layers receives inputs and feeds them to hidden layers. The hidden layers process this input and then forward the outcome to the output layer. In the domain of loan lending, Artificial neural networks have received significant attention for credit scoring and loan default prediction.

Ming-Chun Tsai et al. [34] proposed a Data Envelopment Analysis Discriminant Analysis (DEA-DA) and neural networks based consumer default prediction model. The model performed better compared to their logistic regression, and DA. The neural network reached an accuracy of 98.55% with a hit rate of 85.1%. Amira Kamil Ibrahim Hassan et al. [35] proposed an ensemble neural networks approach for consumer loan default prediction. A two-layer feed forward network is used with three training algorithms. With only 1000 cases in the data, the analysis is limited in terms of data, yet the proposed model has merit, comparing nine models. Selçuk Bayraci et al. [36] proposed a deep learning based loan default prediction model, using deep neural networks alongside logistic regression, naïve bias, and SVM classifier. They conclude that neural networks improve in performance with larger datasets and perform better on more complex datasets. The deep neural network architecture used in the paper gives a score of 85.9% on one of the datasets.

Table 1. STATE OF THE ART PERFORMANCE OF THE EXISTING MODELS ON DIFFERENT FINANCIAL DATASETS.

Study	Detection approach	Accuracy
Kemalbay, G. et al. [13]	2-step Logistic Regression	91.1 %
Agbemava, E. et al. [17]	Logistic Regression	86 %
Moula, F.E. et al. [20]	SVM	85.32 %
Shao, Y.H. et al. [21]	NHSVM	90.68 %
Kim, H.S. et al. [30]	SVM	66.16 %
Eweoya, I.O. et al. [24]	SVM	81.3 %
Zekic-Susac, M. et al. [27]	Decision Tree	74 %
Xia, Y. et al. [29]	Decision Tree + XGBoost	84.65 %
Zhang, D. et al. [30]	VBDTM	81.64 %
		91.67 %
Zhu, L. et al.[31]	Random Forest	98%
Tsai M. C. et al. [34]	NN	98.2%
Bayraci, S et al. [34]	DNN	85.9%

DATASET

The dataset "Loan Default Prediction – Imperial College London" is taken from a Kaggle competition [37]. The competition was uploaded by Imperial College of London, and data has been divided into two sets - train and test data. The train data contains a total of 105471 instances and 771 features, and the test data has 210944 instances. All the feature names have been renamed and the target variable is feature, named as 'target'. Target variable contains a value which defines the amount of loss the lender experienced, normalized between 0 and 100, with 0 denoting no loss, i.e. no default and any value above 0 signifies that the user has defaulted on their loan. The dataset also has missing values that are filled accordingly to retain all the available data. Some features are categorical, for which one-hot encoding is performed. Furthermore, any sort of time dependency is absent in the dataset.

METHODOLOGY

In the paper, we have tested multiple machine learning approaches to achieve loan default classification. In order to achieve so, the following steps have been performed:

1. Data cleaning and standardization

2. Correlational analysis
3. Training machine learning models for Support Vector Machines (SVMs), K-Nearest Neighbors, Random Forest, Neural Network
4. Balancing data and train machine learning models
5. Ensemble Learning.

We have used Receiver Operating Characteristic Curve (ROC AUC) score as our accuracy parameter. This score is particularly used for binary and multiclass classification problems and provides the true performance metric of a model in imbalanced datasets. The dataset is highly imbalanced and so it is necessary to use ROC AUC for testing the relative performance of the machine learning models on both the balanced and imbalanced data.

Data cleaning and standardization

Since there is no semantic information available for the features 771 features, empty and none values have to be critically handled. Over the entire dataset, empty and none values are filled with 0 to make use of all the available data for analysis. An alternate approach is of dropping these instances with none values, however this results in loss of data. For columns with data type as string, i.e. categorical features, these were dropped as one-hot encoding for such a large dataset is computationally challenging and the results were also not effective due to a large number of classes in these attributes. Once the data was preprocessed according to the previously stated steps, all the features were scaled using scikit MinMaxScaler. This kind of normalization allows for faster convergence of models and better performance on the dataset. The dataset is now ready for correlation analysis.

Correlational Analysis

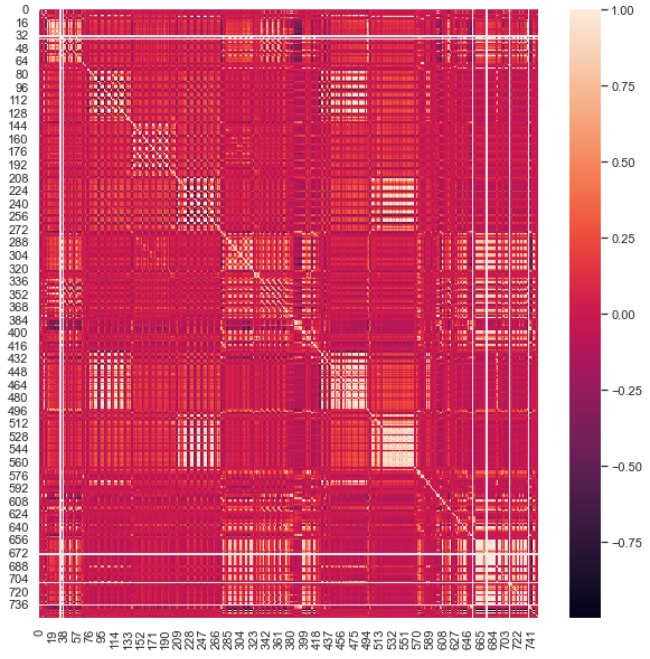

Fig. 1: Correlation Matrix

After standardization of data, the features were subjected to correlational analysis using the pandas library. The plot in Fig. 1. shows the heatmap of the correlation matrix of the features of the dataset. As seen in the correlation matrix of the features, no feature has particularly high correlation with the target variable. Therefore, all the features are used for the initial analysis and initial machine learning training using the preprocessed data.

SVM Model Training

Support Vector Machines are used as the first model for training and prediction The Fig. 2. shows the confusion matrix of its predicted value on validation split of 30% of the train data. For testing, 70% of the train data is used. An interesting thing to note here is that the values for True Negatives and False Negatives are 0, that could signify either that the data has overfitted or there is an imbalance between the positive and negative classifying values.

To further understand the impact of this, we also trained Random Forest and KNN models. Support Vector Machines gave a (ROC AUC) score of 0.5 which is very low. A noteworthy fact about the SVM Classifier is that the training time of the SVM classifier was significantly higher as compared to the other models. The hyperparameters were tuned, with the C value of 2 giving the best results with moderately high train time.

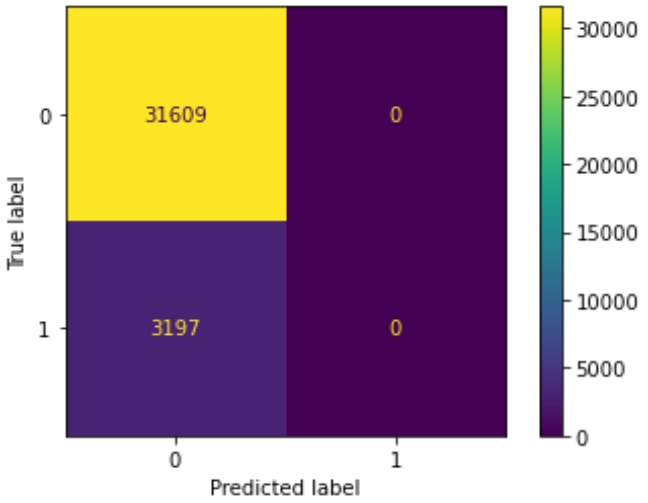

Fig. 2: SVM Classifier Confusion Matrix

Random Forest Model Training

Random Forests Classifier was trained as the next model on 70% of the train data, with the other 30% used for validation. The model reached a ROC AUC accuracy of 0.583, which is an improvement over SVM, yet not very good. A difference in the prediction of False Negatives and False Positives from SVMs was noticed here. Also signifying that Random Forests performed better in identifying the negative cases. A slight increase in ROC AUC of 0.08 is also indication of this improvement.

However, the performance is still not acceptable, and the imbalanced data is not allowing the model to fit properly. Random Forest Classifier is better suited for such imbalanced datasets, as compared to SVM although the current performance of the model is not very effective. Like the SVM classifier, hyperparameter tuning did not improve the performance of the model significantly. The Confusion Matrix for the validation data is shown in Fig. 3. The figure depicts a strong bias for the majority class with extremely poor performance for minority class.

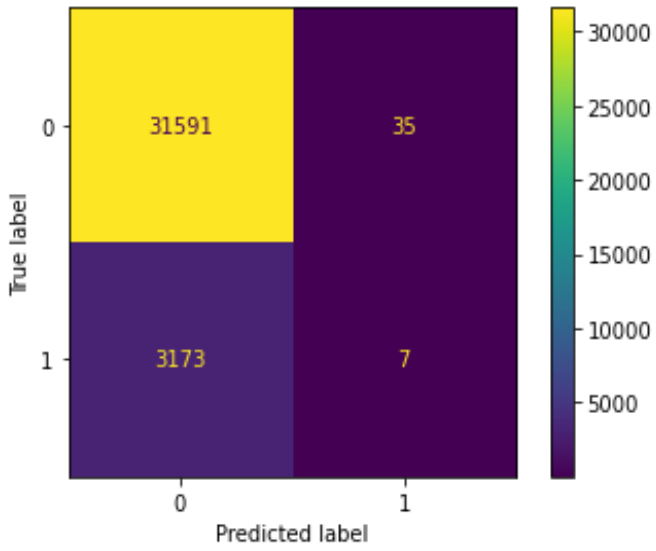

Fig. 3: Random Forest Confusion Matrix

KNN Model Training

The K-Nearest Neighbors model was also trained on 70% of the train data and the model achieved a ROC AUC accuracy of 0.5 that is equivalent to that of the SVM Classifier. The plot showing the confusion matrix in Fig. 4 shows the highest number of false negatives and true negatives. Therefore, we can conclude that KNN performed the best among Random Forest and SVMs to detect negative cases.

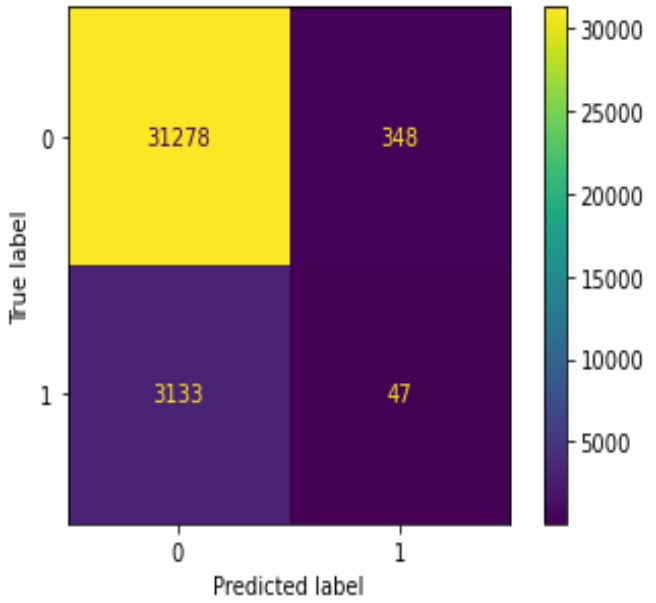

Fig. 4: KNN Confusion Matrix

Neural Network Model

The neural network architecture is demonstrated in the Fig. 5. The relu activation function is used for hidden layers with sigmoid for the output layer. The optimizer is Adam with a learning rate of 0.01. Early stopping is used with minimum delta of 0.0001 and the model is trained for 100 epochs and batch normalization is used with batch size of 500. The model is trained on 70% of the train data. The 30% of the test data is used for validation of the model.

The confusion matrix of the neural network model on the validation data is shown in Fig. 6. It is evident that the model performs poorly with only predicting Positive. This is because the dataset is imbalanced, and the model is overfit to the positive class.

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 751)               564752
_________________________________________________________________
dense_1 (Dense)              (None, 512)               385024
_________________________________________________________________
dense_2 (Dense)              (None, 256)               131328
_________________________________________________________________
dense_3 (Dense)              (None, 128)               32896
_________________________________________________________________
dense_4 (Dense)              (None, 64)                8256
_________________________________________________________________
dense_5 (Dense)              (None, 32)                2080
_________________________________________________________________
dense_6 (Dense)              (None, 1)                 33
=================================================================
Total params: 1,124,369
Trainable params: 1,124,369
Non-trainable params: 0
_________________________________________________________________
```

Fig. 5: Neural Network Architecture

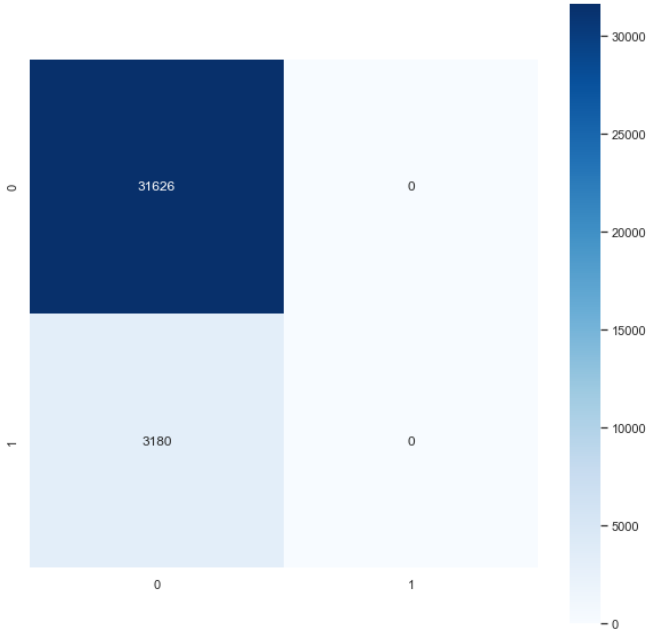

Fig. 6: Neural Network Confusion Matrix

Ensemble Stacking Classifier

Table 2. ENSEMBLE LEARNING MODELS INDIVIDUAL ACCURACY

Detection approach	Individual Accuracy
Logistic Regression	90.8 %
K-Nearest Neighbor	90.0%
Decision Tree	82.4%
Gaussian Naïve Bayes	61.2%

This classifier consists of two levels of stacking classifier, for level 0, four separate classifiers were trained individually.

1. Logistic Regression
2. K-Nearest Neighbor
3. Decision Tree
4. Gaussian Naïve Bayes

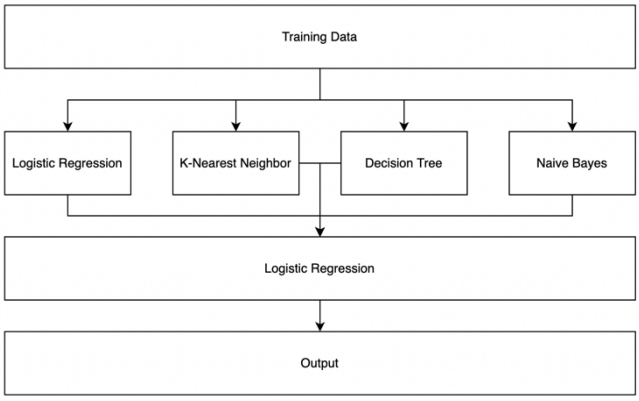

Fig. 7: Stacking Ensemble Learning Architecture

The architecture of the ensemble learning model in shown in Fig. 7. The model is trained on 70% of the train data, rest 30% is used for validation of the model. The confusion matrix on the validation data is shown in Fig. 8. The performance is extremely poor for an ensemble model due to the imbalanced nature of the dataset.

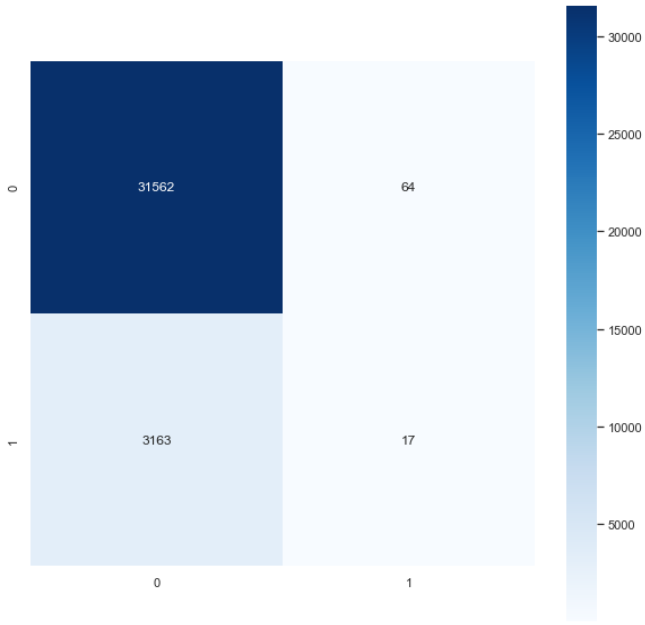

Fig. 8: Stacking Ensemble Model Confusion matrix

Imbalanced Dataset

The poor performance of the models is due to the extremely imbalanced nature of the dataset. The histogram of the classes in the dataset is shown in Fig. 9, demonstrating the high imbalance ratio present in the dataset. The approach must be modified to tailor the training for such datasets.

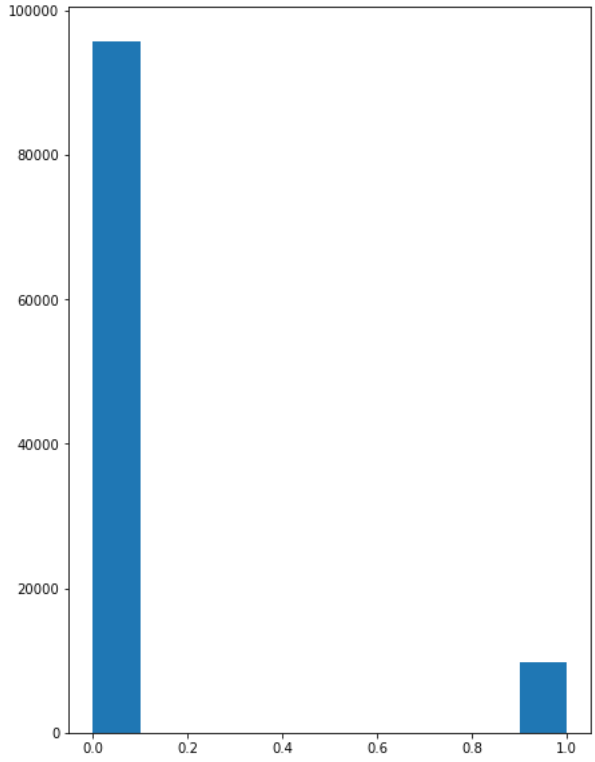

Fig. 9: Histogram of Imbalanced Dataset

Balancing the Dataset

The challenge with imbalanced data is that most machine learning models try to minimize the error rate. This allows them to have high accuracies in the traditional sense, however, they lack in focusing on the minority classes. This can sometimes lead to models just producing one value or completely ignoring the minority class. For real-world applications, especially loan default prediction, imbalanced data is fairly common with extremely high imbalance rations. Figure {} demonstrates how imbalanced the data is, with default instances approximately 1/10 of the non-default instances. This is the primary reason for the machine learning models to perform poorly. In order to improve the performance of the models, the dataset must be balanced. There are three major approaches of dealing with an imbalanced dataset:

1. **Over-Sampling:** This method includes increasing the instances of the minority class to the same number as the majority class(es) by creating synthetic datapoints or duplicating existing ones. The advantage of this approach is that we can use all the available data at our disposal, however it can lead to overfitting of the data.
2. **Under-Sampling:** This method includes decreasing the instances of the majority class to the same number as that the minority class(es). This is achieved by taking random samples from the majority class of the same size of the minority class. The major disadvantage of this approach is that a lot of data may be lost during sampling, that may result in underfitting of the data.

3. **Bagging:** Bagging is commonly used method to deal with imbalanced data. It includes training multiple machine learning models using randomly selected samples of the dataset, mitigating the imbalanced nature of the data.

Under-sampling

Due to the high imbalance ratio of the dataset, the amount of data loss due to under-sampling is considerably high. Due to this, the under-sampled data provides extremely unreliable and volatile results, with the ROC AUC scores of KNN classifier very low. This demonstrates that the KNN model underfits the data due to under-sampling.

Over-sampling

Over-sampling the data resulted in much improved results. The histogram of the dataset classes after over-sampling can be seen in Fig. 10. The dataset is now balanced and a KNN classifier is trained on this balanced data.

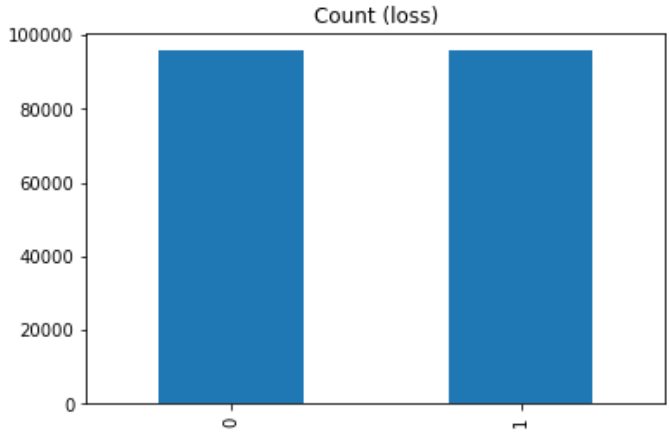

Fig. 10: Histogram of over-sampled data

The Confusion matrix of the KNN model on validation data can be seen in Fig. 11. The improvement in true negatives is clearly evident and the model performs considerably well. The performance improvement is yet another evidence of poor performance of classical machine learning model's deficiency to perform well on imbalanced data.

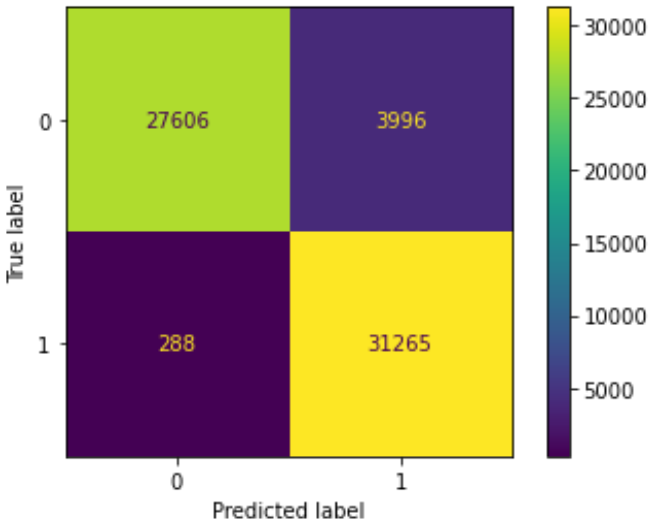

Fig. 11: Confusion matrix for KNN with over-sampled data

Ensemble Bagging Classifier

For Bagging, we have used Balanced Bagging Classifier model, allowing us to resample each subset of data used for training each model. The Confusion Matrix for the model is in Fig. 12. It is noteworthy that it performs significantly better in True Positives for defaulter's class as compared to all the models discussed previously that are trained on imbalanced data.

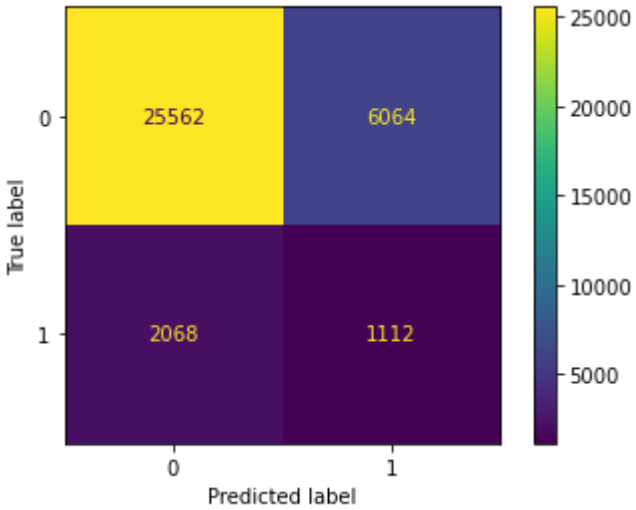

Fig. 12: Balanced Bagging Classifier Confusion Matrix

Ensemble Stacking Classifier on Over-Sampled Data

The architecture of the ensemble learning model in shown in Fig. 7 is now trained on the balanced data. The confusion matrix on the validation data is shown in Fig. 13. The performance improvement is drastic, as seen by the figure, especially in minority class.

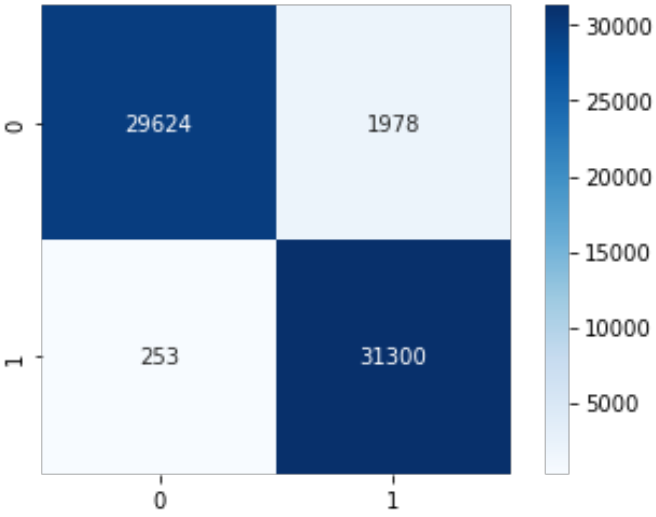

Fig. 13: Ensemble Stacking classifier confusion matrix on over- sampled data

Table 2. ENSEMBLE LEARNING MODELS INDIVIDUAL ACCURACY.

Detection approach	Individual Ensemble Accuracy
Logistic Regression	66.1 %
K-Nearest Neighbor	82.9%
Decision Tree	93.4%
Gaussian Naïve Bayes	60.5%

RESULTS

Amongst the classical classifiers SVM, Random Forest and KNN, even though KNN had better performance for true negatives, the performance improvements are insignificant, and all three models perform poorly on the imbalanced data.

The Balanced Bagging Classifier performed significantly better with and ROC AUC score of 0.636 which is still not satisfactory. For resampling, it was noted that under-sampling the dataset provided poor and highly random results, showing that it is an unsuitable technique for the chosen dataset. On the contrary, over-sampling provided much more stable results, with KNN classifier model giving an ROC AUC score of 52.4, which is marginally better than that on the imbalanced dataset. Stacking Classifier performed best with the balanced dataset, due to its ensemble nature of learning. With an ROC AUC score of 0.704, it performed well. The plot of the ROC AUC scores of the three models on balanced dataset is shown in Fig. 14.

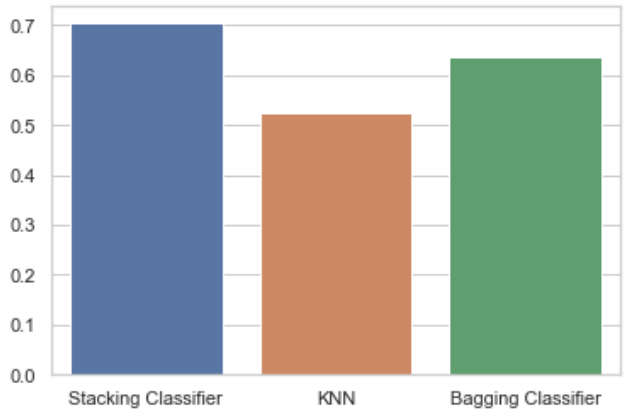

Fig. 14: ROC AUC of different models

DISCUSSION AND CONCLUSION

To conclude, Ensemble stacking classifier performed the best with AUC value of 0.704. Conventional Machine learning models struggle with imbalanced dataset, especially with a high imbalance ratio. Financial data, like credit default detection are primarily imbalanced and therefore, other methods are needed to improve the performance in these sectors. The existing works are promising, yet further work is needed to keep up with the ever-changing needs of the finance world in the current economic scenario. With the increase in instability in the financial sector, with the aftermath global pandemic and the hybrid post-pandemic era, financial institutions are suffering. A more robust credit analysis and default prediction system can aid the economy and support the economic pilers of the economy that are lenders and banks.

FUTURE WORK

Since Ensemble models performed better than singular models, more combinations of ensemble models should be used further for research. With the sheer number of features, more robust feature selection can be incorporated to improve the training process and performance of the models. Use of Neural Networks alongside Ensemble learning could also be explored in future works. In order to solve the issue of universal credit prediction, that is unbiased, yet accurate, extensive work is needed in the domain. Incorporating micro-loans into the analysis increases the complexity of any analysis, due to their impact on smaller economies. Furthermore, taking into consideration the recent unstable market, the relevance of credit is increasing exponentially. With the post-pandemic consumer demand and spending still on the increasing end, the credit system must be a primary focus of research.

ACKNOWLEDGEMENT

We would like to thank Dr Robin Gras for is support and guidance for this project. We would also like to thank Kaggle and Imperial College London for the dataset. Special thanks to University of Windsor for providing the resources for the successful completion of the project.

REFERENCES

[1] C. Yeh, C. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients," Expert Systems with Applications, vol. 36, no. 2, pp. 2473–2480, 2009.

[2] D. J. Hand, W. E. Henley, "Statistical classification methods in consumer credit scoring: a review," Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 160, no. 3, pp. 523–541, 1997.

[3] E. Rosenberg, A. Gleit, "Quantitative methods in credit management: a survey," Operations research, vol. 42, no. 4, pp. 589–613, 1994.

[4] Girija V Attigeri, MM Pai, and Radhika M Pai, "Credit risk assessment using machine learning algorithms," Advanced Science Letters, vol. 23, no. 4, pp. 3649– 3653, 2017.

[5] White, "Credit card debt in the U.S. hits all-time high of $930 billion—here's how to tackle yours with a balance transfer," cnbc.com. https://www.cnbc.com/select/us-credit-card-debt-hits-all-time-high/ (accessed August 5, 2022)

[6] "Delinquency Rate on Single-Family Residential Mortgages, Booked in Domestic Offices, All Commercial Banks [DRSFRMACBS]," Board of Governors of the Federal Reserve System (US), retrieved from FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/DRSFRMACBS.

(accessed August 5, 20220.

[7] S. Arya, C. Eckel, C. Wichman, "Anatomy of the credit score," Journal of Economic Behavior & Organization, vol. *95*, pp.175-185, 2013.

[8] H. A. Abdou, J. Pointon, "Credit scoring, statistical techniques and evaluation criteria: a review of the literature," Intelligent systems in accounting, finance and management, vol. 18(2-3), pp. 59-88, 2011.

[9] U. Aslam, H. I. T. Aziz, A. Sohail, N.K. Batcha, "An empirical study on loan default prediction models," Journal of Computational and Theoretical Nanoscience, vol. 16(8), pp. 3483-3488, 2019.

[10] L. Zhou, H. Wang, "Loan default prediction on large imbalanced data using random forests," TELKOMNIKA Indonesian Journal of Electrical Engineering, vol. 10(6), pp. 1519-1525, 2012.

[11] A.K.I. Hassan, A. Abraham, "Modeling consumer loan default prediction using ensemble neural networks," In 2013

International Conference on Computing, Electrical and Electronic Engineering (ICCEEE), IEEE, pp. 719-724, 2013.

[12] B. Baesens, D. Roesch, H. Scheule, "Credit risk analytics: Measurement techniques, applications, and examples in SAS," John Wiley & Sons, 2016.

[13] G. Kemalbay, Ö.B. Korkmazoğlu, "Categorical principal component logistic regression: a case study for housing loan approval," Procedia-Social and Behavioral Sciences, 109, pp.730-736, 2014.

[14] S.Y. Sohn, D.H. Kim, J.H. Yoon, "Technology credit scoring model with fuzzy logistic regression," Applied Soft Computing, 43, pp.150-158, 2016.

[15] A.C. Bahnsen, D. Aouada, B. Ottersten, "Example-dependent cost-sensitive logistic regression for credit scoring," In 2014 13th International conference on machine learning and applications (pp. 263-269). IEEE, December, 2014.

[16] S.Y. Sohn, D.H. Kim, J.H. Yoon, "Random effects logistic regression model for default prediction of technology credit guarantee fund," European Journal of Operational Research, 183(1), pp.472-478, 2007.

[17] E. Agbemava, I.K. Nyarko, T.C. Adade, A.K. Bediako, "Logistic regression analysis of predictors of loan defaults by customers of non-traditional banks in Ghana," European Scientific Journal, 12(1), 2016.

[18] V. Vapnik, "The nature of statistical learning theory," Springer science & business media, 1999.

[19] A. Mathur, G.M., "Multiclass and binary SVM classification: Implications for training and classification users," IEEE Geoscience and remote sensing letters, 5(2), pp.241-245, 2008.

[20] F.E. Moula, C. Guotai, M.Z. Abedin, "Credit default prediction modeling: an application of support vector machine," Risk Management, 19(2), pp.158-187, 2017.

[21] Y.H. Shao, W.J. Chen, N.Y. Deng, "Nonparallel hyperplane support vector machine for binary classification problems," Information Sciences, 263, pp.22-35, 2014.

[22] D.H. Kim, S.Y. Sohn, "Support vector machines for default prediction of SMEs based on technology credit," European Journal of Operational Research, 201(3), pp.838-846, 2010.

[23] I.O. Eweoya, A.A. Adebiyi, A.A. Azeta, O. Amosu, "Fraud prediction in loan default using support vector machine," In Journal of Physics: Conference Series (Vol. 1299, No. 1, p. 012039). IOP Publishing, August, 2019.

[24] S. Vimala, K.C. Sharmili, "Prediction of loan risk using naive bayes and support vector machine," In *Int Conf Adv Comput Technol (ICACT)* (Vol. 4, No. 2, pp. 110-113), 2018.

[25] B. Baesens, D. Roesch, H. Scheule, "Credit risk analytics: Measurement techniques, applications, and examples in SAS," John Wiley & Sons, 2016.

[26] M. Zekic-Susac, N. Sarlija, M. Bensic, "Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models," In 26th International Conference on Information Technology Interfaces, 2004. (pp. 265-270). IEEE, June, 2004.

[27] S.Y. Sohn, D.H. Kim, "Decision tree-based technology credit scoring for start-up firms: Korean case," Expert Systems with Applications, 39(4), pp.4007-4012, 2012.

[28] P. Bühlmann, "Bagging, boosting and ensemble methods," Handbook of computational statistics (pp. 985-1022). Springer, Berlin, Heidelberg, 2012.

[29] Y. Xia, C. Liu, Y. Li, N. Liu, "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring," Expert systems with applications, 78, pp.225-241, 2017.

[30] D. Zhang, X. Zhou, S.C. Leung, J. Zheng, "Vertical bagging decision trees model for credit scoring," Expert Systems with Applications, 37(12), pp.7838-7843, 2010.

[31] L. Zhu, D. Qiu, D. Ergu, C. Ying, K. Liu, "A study on predicting loan default based on the random forest algorithm," The 7th Int. Conf. on Information Technol. and Quantitative Management (ITQM) 162 pp 503–13, 2019.

[32] N. Ghatasheh, "Business analytics using random forest trees for credit risk prediction: a comparison study," Int. Journal of Advanced Science and Technol. 72 pp 19–30, 2014.

[33] M.A. Mukid, T. Widiharih, A. Rusgiyono, A. Prahutama, "Credit scoring analysis using weighted k nearest neighbor," In Journal of Physics: Conference Series (Vol. 1025, No. 1, p. 012114). IOP Publishing, May, 2018.

[34] M.C. Tsai, S.P. Lin, C.C. Cheng, Y.P. Lin, "The consumer loan default predicting model–An application of DEA–DA and neural network," Expert Systems with applications, 36(9), pp.11682-11690, 2009.

[35] A.K.I. Hassan, A. Abraham, "Modeling consumer loan default prediction using ensemble neural networks," In 2013 International Conference on Computing, Electrical and Electronic Engineering (ICCEEE) (pp. 719-724). IEEE, 2013.

[36] S. Bayraci, O. Susuz, "A Deep Neural Network (DNN) based classification model in application to loan default prediction," Theoretical and Applied Economics, 4(621), pp.75-84, 2019.

[37] Imperial College London, "Loan Default Prediction - Imperial College London", kaggle, 2014.