Credit Card Approval, Credit Default Prediction and Fraud Detection Using Supervised Ensemble Stacking Techniques – A Comparative Study

Abhinav Thapa
M.Sc. Data Analytics
Department of Computing
National College of Ireland
Dublin, Ireland

***Abstract*— Credit card can be defined as the physical manifestation of the credit value, similar to that which is offered by any local currency, but without the hassle of actually physically carrying money for spending. It helps both the customers and businesses to track and manage transactions with ease while speeding up the transaction time for each customer waiting. Thanks to contactless payments and mobile app payments, credit card usage has soared. With its increased usage worldwide, it is critical to keep it safe and secure, to protect the users from frauds and misuse. Despite the availability of countless research on the subject of credit card fraud, approval and default prediction, an elaborate application of stacked learning model is lacking. This study aims to determine whether the accuracy of credit card approval prediction, credit default prediction and fraud detection be improved using Stacking of supervised machine learning models and to find the best performing model among them. Our analysis concluded with Logistic regression as the best meta-classifier (99.5% accuracy) followed by Decision Trees while Support Vector Classifiers and K-Nearest Neighbors turned out to be very computationally expensive when used as base learners in a Stacked Ensemble model.**

Keywords—Approval prediction, Default prediction, Fraud detection, Stacked Ensemble Machine Learning, RF, XGB, DL, SVC, NB, KNN, EDA, SMOTE, Random Sampling.

I. INTRODUCTION

The concept of using a card for purchases was first mentioned in 1887 by an American author, journalist and activist, Edward Bellamy in his utopian novel Looking Backward, but no one could have known at the time how much of a financial significance it would hold in the future. Today, owning a credit card is synonymous with higher purchasing power, priority perks, and even higher social status. Over the last few decades, with the exponential rise of e-commerce and the associated emerging technologies, the spending habits of customers have increased which has ultimately led to the boom of credit cards usage across the globe. In 2022, there are 1.06 billion credit cards in use in America, and 2.8 billion throughout the world [21] .

However, with all the benefits and convenience it has to offer to its customers, it also presents a huge liability for banks and financial institutions as it is susceptible to frauds and defaults. Left unchecked, the whole financial institution could come to its knees over a short period of months. Hence, it's imperative to provide such services to only appropriate customers and to perpetually employ advanced technologies to secure the organization against frauds, defaults and misuse. Artificial Intelligence and Machine Learning has proven to be the most promising approach to anticipate the suitability of a prospective customer based on the financial history and personal profile, default prediction based on the customer's payment history and fraud detection using anomalous spending behaviour.

Despite the far-reaching applications and capabilities of AIML, achieving a machine learning model ideal for a certain use case is not always feasible. Past research in the field has shown that using supervised learning methods such as Logistic Regression, Boosting (XGB, AdaBoost, GBM), Bagging (RF), Deep Learning (ANN, NNET) and Stacked model algorithms yield results that satisfy industry standards, however, elaborate literature on the Stacked model comparison is lacking.

This paper seeks to understand the effect of using the Stacked Ensemble method of using supervised learners (LR, SVM, RF, NB, DT) as base classifiers and stacking them using a meta-classifier (LR, SVM, DT) to evaluate and possibly achieve a higher metric using one of the suitable Oversampling techniques such as SMOTE, and Cross-validation technique such as k-fold. Accuracy will be used as metrics to gauge the effectiveness of each model and achieve the best performing model for each use case i.e. Credit approval prediction, Default prediction and Fraud detection.

II. RELATED WORK

The following literature review elaborates on the analysis and findings from other researches performed on the concerned topics.

Extensive research on Credit card fraud detection has already been done. But most of the literature explores the high-class imbalance nature of the data or learns the customer behaviour to predict anomalous transactions. Same could be said for credit default prediction problem as well, however, credit approval prediction has not been explored to such extent before. Unavailability of real datasets due to the sensitive nature of information contained and the highly imbalanced nature of the data have posed hurdles for researchers from literature point of view.

A. Credit Fraud Detection Literature Review

Veigas, K.C., Regulagadda, D.S., Kokatnoor. S.A. (2021) [1], explored a Stacked model by combining standard classifiers (KNN and SVM) with Multi-layer Perceptron (MLP) on SMOTE and GAN augmented dataset for fraud detection. Although near-perfect accuracy and increased F1-score were achieved, it was evident that the proposed model lacked computational simplicity and performance speed. Also, further comparison among multiple instances of Stacked models could have provided valuable insights.

Gupta, S. (2016) [2], attempts comparison of the performance of Deep Learning (DL) with RF, GBM and GLM algorithms used with sampling techniques (Hybrid, Over, Under, SMOTE and ROSE) for fraud detection. DL algorithms performed best with the highest Recall but with lower Precision. Using sampling techniques did help in improving AUC and precision for GLM while best F-score was achieved for the model trained using ROSE sampling. The study used H2O library and suggested that the use of advanced DL libraries such as Torch, Theano, Tensorflow, Keras etc. could help to provide better performance overall.

Brennan et al. (2012) [3], dwells on the problem of high-class imbalance on the fraud detection datasets. Two approaches were proposed to treat them: Algorithmic level and Data level. At the algorithmic level, cost-sensitive learning was performed while at the data level approach, artificial re-sampling of the minority/majority classes using SMOTE was performed. As a result, the data approach concluded that oversampling of minority classes showed better performance than under-sampling of majority classes. However, algorithmic approach concluded that the RF learner was best among Naïve Bayes, ID3, C4.5, KNN and RIPPER.

Abdulla et al. (2015) [4], used a hybrid approach that focused on pre-processing and removed anonymous records, then used a genetic algorithm (GA-KNN) for feature selection, and finally used SVM for classification. The proposed model did achieve a better accuracy, however, no evaluation for the efficacy of the hybrid approach was provided.

Chalwadi et al. (2021) [5], attempts to solve the problem using Neural Network Multi-Layer Perceptron classification with upto 2 hidden layers on the PCA components of the dataset along with ADASYN sampling. Accuracy achieved was high and comparative to the ML models (RF, Boosting, NB and DT), however, a comparatively poor F1 score could be achieved. A comparative study nonetheless, but further evaluation to ascertain the suitability of the models with the problem statement was lacking.

Patil, T. (2021) [6], used a hybrid approach for Fraud detection utilizing CT-GAN data modelling technique and SelectKBest feature selection and built 3 models using RandomForest (RF), Logistic Regression and Extreme Gradient Boosting (XGB) algorithms for evaluation and achieved perfect recall and F1-scores with RF model, however, the limitations and implications of using an augmented dataset were not explored in detail. Additionally, exploring the approach on using categorical data could have increased its reliability.

Soleymanzadeh, R.; Aljasim, M.; Qadeer, M.W.; Kashef, R. (2022) [7], discussed the application of stacked ensemble classifier to detect cyberattacks in IOTs with high performance on three different datasets: credit card, NSL-KDD and UNSW datasets. The stacked ensemble technique with two approaches were used, i.e., one by combining poor learners and, the other by combining strong learners. The stacked ensemble of poor base learners generally performed better than the stacked ensemble of strong base learners and also provided a better computational speed for all the datasets. The paper generalizes effectively on the application of the Stacked Ensemble of poor vs strong base learners for cyberattacks and credit data, however an in-depth comparative analysis is imperative to realise its feasibility and reliability.

B. Default Prediction Literature Review

Egan, C. (2021) [8], focuses on explainable Artificial Intelligence aspects of the models used for the problem of Credit default prediction. It's a notable paper as it elaborates on the explainability of the black-box models rather than only on the performance, by extracting counterfactual extraction algorithms to make them more understandable. Given that the paper utilized the counterfactual extraction methods to arrive at translatable rules that can be understood easily, enough clarity on its application on different datasets to gain a complete understanding of the algorithms is missing. Also, evaluation of this method on more generalized use-cases might have helped.

Yeh, I. C., and Lien, C. H. (2009) [9], devised an interesting approach to conduct a comparative analysis of six data mining techniques to evaluate the predictive accuracy for probability of default rather than the standard binary-label classification using Sorting Smoothing Method. Even though the relative differences in errors were little, there were substantial differences in probability area ratio due to which the Artificial Neural Network (ANN) turned out as the top scoring model with R^2 (0.96), regression intercept (0.01) and coefficient (0.99). The paper provided a thorough analysis of the probability prediction and suggested this approach for classifying clients, however, failed to address the relevant implications associated with transitioning to such a system for Credit companies and users as most of the time having a numerical threshold in real scenarios is not possible for informed decision-making.

Subasi, A. and Cankurt, S. (2019) [10], used SMOTE variations (without, with 100% and with 200%) with seven data mining algorithms for default prediction and found that RF with 200% SMOTE performed best in the field with least errors, improved ROC and highest F-measure. It was emphasized that utilizing simple algorithms with suitable sampling techniques could also provide outstanding results in terms of performance and accuracy. However, comparison with hybrid models based on accuracy and performance was lacking in the report.

Peng, R. (2017) [11], claims that the model accuracy and robustness are both handled efficiently by a 2-layer Stacking Ensemble learner and is able to perform better (Accuracy Training set – 94.66% and 85.31% on Test set) than the tuned individual learners (SVM, RF, ANN, GBDT and Simple Voting). Although a detailed comparison among the individual models against the Stacked Model, a comparative study of the Stacked Model instances and their performances is absent.

Lu, H., Haifeng W., and Sang W. Y. (2017) [12], discussed about the advantages of Online Adaboost learning and Extreme learning machine methods and analysed the accuracy and computational performance with their corresponding offline learning methods along with other algorithms. Online methods maintained their accuracies but were significantly more efficient in terms of computational speed and more reliable and flexible for use cases with transient datasets (as in credit default prediction). Further exploration on the stability and robustness of the models would have shed some more light on the feasibility and reliability of such an approach for Default prediction.

Zurada J. (2010) [13], compared the accuracy scores for eight models: logistic regression (LR), neural network (NN), radial basis function neural network (RBFNN), support vector machine (SVM), case-based reasoning (CBR) and three Decision Tree (DT) models, by applying 10-fold cross-validation and repeating the experiment 10 times. The DT models provided the best accuracy scores and they proved to be much more interpretable and practical for explaining the results. Incorporating further analysis by including Random Forest (RF) model could have enriched the comparative study and highlighted the reason for potential popularity of the RF models that could have been more effective.

C. *Credit Approval Prediction Literature Review*

Sakprasat S. and Sinclair M. C. (2007) [14], used the multiple approaches of Genetic Programming (GP) for addressing the automatic credit approval use-case. Though the accuracy was not much higher when compared to pre-processing approach on testing set, the GP approach was able to handle the modelling phase with the data containing missing values and still provide a good generalized prediction. However, the proposed model is still not immune to the deviations caused by redundant data, misleading data (outliers) and data size (not suitable for large datasets). A comprehensive study on the appropriate data size permissible for GP approach and a comparative study with other non-GP approaches could enrich the analysis provided by the authors.

Abakarim Y., Lahby M. and Attioui A. (2018) [15], explored the Deep Learning Neural Network with auto-encoders for classification of applicants and compared the proposed model with several typical classification models. It was evident that the proposed model out-performed even the most effective SVM model. Although comparative study was provided with typical models, application of Ensemble techniques were not included which could have been more illuminating. Hyper-parameter tuning on the Neural Network could further improve the efficiency of the proposed model.

Pristyanto Y., Adi S. and Sunyoto A. (2019) [16], elaborates on the effects of correlation-based and information gain-based feature selection techniques on the credit approval classification problem. Almost always the accuracy increased for both approaches in KNN, SVM, ANN and DT models, except when the information gain-based feature selection was applied to DT, the accuracy did not increase. Further evaluation of diverse datasets and elaborate comparison could provide valuable insights into the reliability of the proposed methods.

Lusinga, Moses, et al. (2021) [17], provided a comprehensive analysis of the statistical classification and machine learning techniques on Home Credit, Xente and Super Lender datasets from developing states (Uganda, Nigeria and US) while also evaluating the interpretability and decision-making (using SHAP) for the proposed models to make them more practical and reliable. It was observed that almost all ML methods were more effective than the almost the statistical methods except in some cases for ANN due to lack of data and high model complexity. Finally, XGBoost when used with SHAP for feature selection yielded the best outcome and made the model as much feasible as accurate for prediction. The study made some interesting inferences, albeit on a limited set of datasets. Also, further analysis of the other techniques from popular Deep Learning could provide a complete spectrum of the proposed models performance and reliability.

III. Methodology

Knowledge Discovery in Databases (KDD) [18] is one of the most reliable and popular data mining methodology (Fig. 1) for extracting valuable knowledge from vast and complex datasets. It is an iterative and adaptive technique that not only utilizes the fundamentals of the data mining approach but also allows to combine the business/domain knowledge in the analysis. Hence, it was chosen for the purpose of this project.

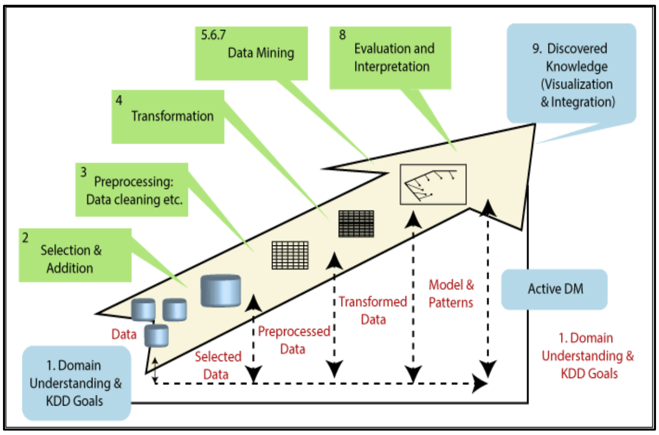
5.6.7
Data Mining
8
Evaluation and Interpretation
9. Discovered Knowledge (Visualization & Integration)
4
Transformation
3
Preprocessing: Data cleaning etc.
2
Selection & Addition
Data
Selected Data
Preprocessed Data
Transformed Data
Model & Patterns
Active DM
1. Domain Understanding & KDD Goals
1. Domain Understanding & KDD Goals

Fig. 1: KDD process of data mining [18]

A. *Domain Understanding and KDD goals*

Economic stability of a country is one of the most important pillars of democracy in the world right now. Once, the financial system starts to shake the world as we know it could change dramatically for us. Hence, it is imperative to employ the necessary resources to keep updating our policies and technologies to address this serious issue of financial credit related problems as it is vulnerable to cyberattacks, theft, misuse, misrepresentation and frauds.

This study aims to analyse this problem by analysing Credit Approval, Credit Default, and Credit Fraud Detection use-cases using Stacked Ensemble Learning and provide critical acumen from the chosen datasets.

B. *Dataset Selection and Description*

Three datasets were chosen from the public domain and used for each of the use cases as mentioned below:

1. *Credit Card Approval Prediction Dataset*

This dataset from Kaggle consisted of two tables – one for applicants record and the other for credit record. The “application_record.csv” (**438557 x 18**) contained the demographics and family data (***predictors*** such as Gender, Car ownership, Total income, Housing type, Occupation, No. of family members etc.) of each applicant required to gauge their eligibility to own a credit card, while the “credit_record.csv” (1048575 x 3) includes data for past payment/default (Balance and ***response variable***- **Status**) for a given customer.

As expected, the dataset is highly imbalanced and contains duplicates due to which it may need extensive EDA and transformations before analysis can be considered.

Source: *https://www.kaggle.com/rikdifos/credit-card-approval-prediction*

2. *Credit Default Prediction Dataset*

This dataset from UCI repository consists of **30000** records and **24** features for default payments in Taiwan in

2015. The dataset contains not only the demographic info of clients but also their credit payment history of past months as features which can be used to predict whether they will default their payment or not (response).

Source: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

3. *Credit Card Fraud Detection Data*

Due to the increased confidentiality and regulatory compliance requirements, it is very difficult to obtain a real dataset for credit card transactions. Due to which the dataset found in Kaggle has principal components as its features that will be used for our analysis in order to secure sensitive customers data.

This dataset contains credit card transactions by European customers in September 2013. It has **284807** transactions and **31** features. We don't have much details of the 28 features (V1-V28) but we have Time and Amount for each transaction and the response variable i.e. the Class 0 or 1 stating whether the transaction was a fraudulent one or not. Here we have only 498 fraudulent records out of almost a quarter of a million records i.e. 0.198 % of the minority class (Frauds). This huge class imbalance will need to be treated appropriately before any analysis for study can be considered.

Source: https://www.kaggle.com/mlg-ulb/creditcardfraud

C. *Pre-processing and Transformation*

Pre-processing and Exploratory Data Analysis go hand in hand as in this step all of the data irregularities and inconsistencies (missing values) are dealt with using suitable visualizations and transformation techniques, in order to prepare it for Machine learning tasks and make it suitable to draw out critical and relevant knowledge. Following essential steps were taken and important insights noted for each of the datasets for this phase:

1. *Credit Card Approval Prediction Dataset*

- Datasets (credit_record.csv, application_record.csv) summary revealed relatively highly skewed distributions in the numerical features while the categorical features including the response variable (*STATUS*) had high class imbalance.

- High number of missing occupation data (30.6 %) called for feature engineering using imputation in the *OCCUPATION_TYPE* feature by the most popular Occupation in that category, in place of the blank missing places.

- *DAYS_BIRTH* and *DAYS_EMPLOYED* were converted from days units to units in years into *AGE* and *YEARS_EMPLOYED* features.

- Label encoding with numerical labels for non-binary categorical predictors.

- Transformed 8 categories in ***response*** variable – *STATUS* into 3 categories – "*Good_Debt*", "*Neutral_Debt*" and "*Bad_Debt*" for Approval prediction using Binary classification.

- Evaluated total count of each response category and assigned an Approval Status in the feature - *CREDIT_APPROVAL_STATUS* in credit dataset.

- Finally merged both transformed datasets (applications and credit status) on the basis of *ID* feature and feature scaled age and income features as they were not normalized.

- Post feature scaling and Oversampling Minority class (0.48%: Denial of application) using SMOTE, we fixed the problem of class imbalance, however increased data points to be used in a complex Stacked model increased our computational time. So, **we used the unsampled dataset for this use case**. Next, we started with Model building and evaluation of ML algorithms which is explored in detail in the Data Mining phase.

2. *Credit Default Prediction Dataset*

- The excel file was loaded and the first unnecessary row and column *ID* were removed.

- Since the response variable *default payment next month* had fairly enough class balance, we did not perform SMOTE oversampling and considered the whole dataset for evaluation.

- No missing values were found.

3. *Credit Card Fraud Prediction Dataset*

- Due to confidential nature of data (creditcard_fraud.csv), this dataset contained only the principal components (*V1-V28*) extracted from the real features as predictors including *Time* and *Amount* while *Class* (0: Not fraud, 1: Fraud) as the *response* variable with huge class imbalance (0.172% Frauds only).

- Transformed *Time* from seconds into hours and performed EDA for feature importance.

- SMOTE for addressing major class imbalance of the minority class (Not Frauds classes) is very essential as it is a sensitive use case and high risk may be dangerous.

- Random Sampling (0.05%) for reducing total data points in order to relieve some of the problems relating to computational time and feasibility.

D. *Data Mining and Evaluation*

Due to the problem of high- class imbalance, pre-processing of the dataset is necessary in order to expect unbiased predictions. For all the three datasets one of the re-sampling techniques such as Synthetic Minority Oversampling Technique (SMOTE) was applied based on feasibility in terms of computational time and model complexity. For each of the resampled dataset, Supervised learning techniques such as Decision Tree, Logistic Regression, Naïve Bayes, K-Nearest Neighbor, and Support Vector Machine will be applied and the Classification metrics from the cross-validation sets will be evaluated. Ensemble techniques such as Random Forest (RF) and Gradient Boosting (XGB) were also evaluated against the supervised methods. Finally, Stacked Ensemble models will be built to achieve a higher metric as compared to other models. Stacking is an Ensemble Machine Learning technique wherein a meta-learning algorithm learns to combine the predictions from its base learners [6]. A stacking model utilizes the best attributes of the base models and combines them while generalizing the results. It includes Base-Models (level-0 Models) and Meta-Models (level-1 Models). The Base models fit on the training data while the Meta-Model learns the best way to combine the predictions of the base models.

Stacking is designed to improve modelling performance, although it is not guaranteed to result in an improvement in all cases[6]. Achieving a performance improvement will depend upon the choice of base learners and the meta-learners due to which multiple combinations of the aforementioned algorithm shall be evaluated. Base learners could be chosen from the Supervised Learning Classification techniques (LR, SVM, NB, KNN and DT) or Ensemble Techniques (RF, XGB and ADAboost). Meanwhile, the Meta-learners could be chosen from one of the Supervised Learning techniques (LR, SVM, NB, KNN and DT) for combining the predictions from the base learners. Although the complexity of the final model may increase, the purpose for this research is to find out whether further improvement of performance can be achieved using Stacking techniques on the resampled datasets and is it feasible.

1. *Credit Card Approval – Model Building*

S.no.	Model Name	Base Learners	Meta Learners	Wall Time	Accuracy Scores (%)
1	DT	-	-	-	99.44
2	LR	-	-	-	99.58
3	SVC	-	-	-	99.58
4	NB	-	-	-	99.52
5	KNN	-	-	-	99.50
6	RF	-	-	-	99.58
7	Xgb	-	-	-	99.58
8	Stacked_LR1	DT, LR, SVC, NB	LR	2 min 22s	99.5
9	Stacked_LR2	DT, LR, SVC	LR	2 min 23s	99.5
10	Stacked_LR4	SVC, NB	LR	3min 36s	99.5
11	Stacked_LR5	DT, LR, NB	LR	14.7 s	99.5
12	Stacked_DT1	DT, LR, SVC, NB	DT	3min 42s	98.1
13	Stacked_DT2	DT, LR, SVC	DT	3min 44s	98.4
14	Stacked_DT3	DT, LR	DT	10.9s	98.58
15	Stacked_DT4	SVC, LR	DT	3min 38s	98.09
16	Stacked_SVC1	LR, DT, SVC, NB	SVC	3min 44s	99.51

Table 1: Approval Data Model Results***

Although Oversampling helped to address the problem of high-class imbalance among *STATUS* categories, **we decided to use the unsampled dataset so as to utilize the present computational capacity of the research infrastructure and achieved the best accuracy scores** (Table 1 models 8,9,10,11,16). Further analysis in future on Stacked Learning algorithms can include use of the extensive datasets for evaluation using higher computational capacity from high-end computers.

2. *Credit Default Prediction – Model building*

S.no.	Model Name	Base Learners	Meta Learners	Wall Time	Accuracy Scores (%)
1	DT	-	-	-	73.48
2	LR	-	-	-	81.13
3	SVC	-	-	-	81.82
4	NB	-	-	-	70.18
5	KNN	-	-	-	79.68
6	RF	-	-	-	80.68
7	Xgb	-	-	-	81.83
8	Stacked_LR1	DT, SVC, NB	LR	4min 53s	77.89
9	Stacked_LR2	DT, LR, SVC	LR	5.02s	78.09
10	Stacked_LR3	DT, LR NB	LR	3.62s	78.19
11	Stacked_LR4	DT, NB	LR	4.25s	77.88
12	Stacked_DT1	LR, SVC, NB	DT	3min 21s	67.84
13	Stacked_DT2	LR, NB	DT	2.1s	68.04
15	Stacked_SVC1	DT, LR	SVC	46s	77.87
16	Stacked_SVC2	DT, NB	SVC	47.3s	77.99

Table 2: Default Prediction Model results***

Sampling was not required here as reasonable records were available from both target classes for evaluation.

3. *Credit Card Fraud Detection – Model Building*

S.no.	Model Name	Base Learners	Meta Learners	Wall Time	Accuracy Scores (%)
1	DT	-	-	-	98.53
2	LR	-	-	-	95.9
3	SVC	-	-	-	66.30
4	NB	-	-	-	87.15
5	KNN	-	-		84.3
6	RF	-	-	-	99.37
7	Xgb	-	-	-	99.73
8	Stacked_LR1	DT, SVC, NB	LR	2 min 19s	95.50
9	Stacked_LR2	DT, LR	LR	19.9s	98.54
10	Stacked_LR3	DT, LR, NB	LR	19.4s	98.56
11	Stacked_LR4	DT, NB	LR	18.1 s	98.51
12	Stacked_DT1	LR, SVC, NB	DT	39.9s	95.51
13	Stacked_DT2	LR, NB	DT	6.5s	96.06
14	Stacked_SVC1	DT, LR	SVC	22.8s	98.50
15	Stacked_SVC2	DT, NB	SVC	17s	98.56

Table 3: Credit Fraud detection Model results***

**Reference: LR – Logistic Regression, DT – Decision Trees, SVC – Support vector classifier, NB – Naive Bayes, Xgb – Extreme Gradient boosting, RF – Random Forest, KNN – K Nearest Neighbor algorithms.

Post **Oversampling** to balance class in response variable, and then **random sampling** to reduce computational load for the purposes of this project**, we were able to achieve higher speed but this reduced some of the accuracy scores for the simpler models in comparison to the Stacked Learners** in Table 3.

Due to high computational demand of SVC and KNN models, they were not preferred for this step of stacked learning. Selected models were evaluated with possible combinations of base (LR, DT, SVC, NB) and meta (DT, LR, SVC) learners in the stacked technique and the results were obtained in Table 1, 2 and 3.

E. Implementation

- The datasets for the project topic were sourced from the internet repositories and selected post in-depth analysis of the project idea and relevance.
- Some of the pre-processing and EDA were inspired from earlier research and analysis. This project further analyzed the Stacked Ensemble techniques and its effects by providing a comparative analysis in terms of computational cost and accuracy as novelty for this paper.
- Programming for this research was conducted primarily on **Jupyter Notebook** using Python 3 kernel.

Post analysis results were recorded and the code were stored for each use case separately.

IV. Results

Extensive Exploratory Analysis and Model building evaluation yielded the following results:

A. Credit Approval Use case

1) Logistic Regression and Support Vector Classifier performed the best among the simple models (99.58%).

2) RandomForest and Extreme Gradient Boosting performed almost similarly (99.58%).

3) Among the Stacked Ensemble Learners, Stacked_LRs and *Stacked_SVC1* (Table 1) perfomred the best with 99.5% accuracy.

4) Stacked_DT3 (Table1, Serial-14: 10.9s), *Stacked_LR5* (Table1, Sr.11 : 14.7s) were the fastest while *Stacked_SVC1* (Table 1 Sr.16: 3min 44s) was the slowest in terms of computational speed.

5) Oversampling using SMOTE helped the class imbalance problem by generating minority class cases, however, for larger datasets this exponentially increased computational costs.

B. Credit Default Use cae

1) XGB and *SVC* were the best performers with 81.83 and 81.82% scores while NB was the poorest (70.18%) for this dataset.

2) Stacked_LR2 and *Stacked_LR3* performed almost similarly (78.1%) and were the best among the Stacked learners and among the fastest as well (5.02s and 3.62s).

3) Stacked_LR1 and *Stacked_DT1* were the slowest (4 min 53s and 3 min 21s) and more complex.

C. Credit Card Fraud Detection Use case

1) Random Forest and Extreme Gradient Boosting Technique performed the best among the simpler models (99.37 and 99.73% respectively).

2) SVC performed very poorly (66.3%) compared to other models.

3) Stacked_LR3 (Table3 Model10 :19.4s) and *Stacked_SVC2* (Table3 Model15 :17s) performed the best among the Stacked learners and were relatively faster than others as well.

4) Stacked_DT2 and *Stacked_SVC2* held the fastest record of 6.5s and 17s respectively for this dataset while *Stacked_LR1* was the slowest (2min 19s).

5) Multiple sampling techniques improved the nature of our dataset, however, to make the analysis feasible for Stacked learner analysis, some of the data was left out of analysis.

Analyzing each use case, we recorded the final best performing Stacked Ensemble Learning Models in terms of complexity, speed and accuracy below:

Table No.	**Best Stacked Learners**		
	Model Name and Build	***Accuracy***	***Wall Time***
1*	Stacked_LR5 (lv0: DT, LR, NB; lv1: LR)	99.5	14.7s
1*	Stacked_DT3 (lv0: DT, LR; lv1: DT)	98.58	10.9s
2*	Stacked_LR2 (lv0: DT, LR, SVC; lv1: LR)	78.09	5.02s
2*	Stacked_LR3 (lv0: DT, LR, NB ; lv1: LR)	78.19	3.62s
3*	Stacked_LR3 (lv0: DT, LR, NB; lv1: LR)	98.56	19.4s
3*	Stacked_SVC2 (lv0: DT, NB; lv1: SVC)	98.56	17s

Table 4: Stacked Learners Model results for 3 datasets

V. Conclusions

Stacked Ensemble Learning is the right approach when the necessity for peak performance is essential no matter the complexity and computational effort, however, such algorithms are highly dependent upon dataset size and complexity. While increasing complexity can improve performance, it is not certain with SVC and KNN algorithms that use spatial metrics for prediction. We also observed that SVC and KNN are not preferably suitable for large datasets and complex ensemble techniques as they become very computationally expensive for each increased quantity of base learner.

Simpler models (LR, DT, NB) when used as Base-learners yield better results with Logistic Regression as Meta-Learner with higher speed.

Our in-depth analysis, only halted by computational capacity, completed the research with the inference that Stacked Ensemble Techniques can improve the accuracy of the models compared to other techniques, however, careful evaluation and model selection will help in reducing complexity and, increasing speed and feasibility.

Further analysis with improved infrastructure could explore this limitation and provide further elaborate conclusions.

REFERENCES

[1] Veigas, K.C., Regulagadda, D.S., Kokatnoor. S.A., “Optimized Stacking Ensemble (OSE) for Credit Card Fraud Detection using Synthetic Minority Oversampling Model”, School of Engineering and Technology, CHRIST , 2009, Bangaluru, India .

[2] Gupta, Sapna, “Deep Learning vs. traditional Machine Learning algorithms used in Credit Card Fraud Detection”, 2016, Masters thesis, Dublin, National College of Ireland.

[3] Brennan, P. “A comprehensive survey of methods for overcoming the class imbalance problem in fraud detection”, 2012, 1-107.

[4] Abdulla, N., Rakendu, R., & Varghese, S. M., “A Hybrid Approach to Detect Credit Card Fraud”, 2015, 5(11), 304-314.

[5] Chalwadi, Ketan R., “Classification of Credit Card Fraudulent Transactions using Neural Network and Oversampling Technique”, 2021, Masters thesis, Dublin, National College of Ireland.

[6] Patil, T., “Credit Card Fraud Detection Using Conditional Tabular Generative Adversarial Networks (CT-GAN) and Supervised Machine Learning Techniques”, 2021, Masters thesis, Dublin, National College of Ireland.

[7] Soleymanzadeh, R.; Aljasim, M.; Qadeer, M.W.; Kashef, R. (2022) Cyberattack and Fraud Detection Using Ensemble Stacking. AI 2022, 3, 22–36. https://doi.org/10.3390/ai3010002

[8] Egan, C., “Improving Credit Default Prediction Using Explainable AI”, 2021, Masters thesis, Dublin, National College of Ireland.

[9] Yeh, I-Cheng, and Che-hui Lien. "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients." Expert systems with applications 36.2 (2009): 2473-2480.

[10] Subasi, Abdulhamit, and Selcuk Cankurt. "Prediction of default payment of credit card clients using Data Mining Techniques." 2019 International Engineering Conference (IEC). IEEE, 2019.

[11] Peng, Runze. "Personal Credit Assessment Model Based on Stacking Ensemble Learning Algorithm." Statistics and Application (2017): 411-417.

[12] Lu, Hongya, Haifeng Wang, and Sang Won Yoon. "Real time credit card default classification using adaptive boosting-based online learning algorithm." IIE Annual Conference. Proceedings. Institute of Industrial and Systems Engineers (IISE), 2017.

[13] J. Zurada, "Could Decision Trees Improve the Classification Accuracy and Interpretability of Loan Granting Decisions?," 2010 43rd Hawaii International Conference on System Sciences, 2010, pp. 1-9, doi: 10.1109/HICSS.2010.124.

[14] S. Sakprasat and M. C. Sinclair, "Classification rule mining for automatic credit approval using genetic programming," 2007 IEEE Congress on Evolutionary Computation, 2007, pp. 548-555, doi: 10.1109/CEC.2007.4424518.

[15] Y. Abakarim, M. Lahby and A. Attioui, "Towards An Efficient Real-time Approach To Loan Credit Approval Using Deep Learning," 2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC), 2018, pp. 306-313, doi: 10.1109/ISIVC.2018.8709173.

[16] Y. Pristyanto, S. Adi and A. Sunyoto, "The Effect of Feature Selection on Classification Algorithms in Credit Approval," 2019 International Conference on Information and Communications Technology (ICOIACT), 2019, pp. 451-456, doi: 10.1109/ICOIACT46704.2019.8938523.

[17] Lusinga, Moses, et al. "Investigating Statistical and Machine Learning Techniques to Improve the Credit Approval Process in Developing Countries." 2021 IEEE AFRICON. IEEE, 2021.

[18] Web reference: “KDD- Knowledge Discovery in Databases”; https://www.javatpoint.com/kdd-process-in-data-mining

[19] Web reference: “Credit card statistics 2022: 65+ facts for Europe, UK, and US”; https://blog.spendesk.com/en/credit-card-statistics

[20] Web reference: “Stacking Ensemble Machine Learning With Python”; https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/

[21] Web reference: “SMOTE for Imbalanced Classification with Python”; https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[22] Web reference: “Metrics to Evaluate your Classification Model to take the right decisions” by sumeet51 — July 20, 2021. https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/