**Student Name: Balan Varadharajan**

**Programme: UFCF9Y-60-M-CSCT MASTERS PROJECT 22SEP\_1FT**

**Student ID: 21072115**

**Building a Hybrid Credit scorecard using Machine Learning & Deep Learning with Knowledge Distillation**

**Introduction:**

The Credit Score assessment is a numeric expression estimating people's creditworthiness. Bank use to judge the people based on the credit score, who is eligible for the loan, at what interest rate, and credit limits. The purpose of this research is to use credit scoring techniques to analyze the probabilities of default (PD), and build a credit scorecard to assign credit scores for the existing customer and new customer, using Multiple Machine Learning Algorithms and Deep Learning Techniques and top of this will use the Knowledge Distillation techniques to increase the prediction accuracy.

In recent years, many financial institutions and scholars have studied many different credit scoring models due to the release of many credit-related data in some countries [1]. Risk management plays significant role in the corporate strategy of many huge financial institutions and corporations, which is usually sustainable [2]. On the off chance that there is just default data, it is hard to choose whether there is any bank fraud or related archive misrepresentation [3]. Customer are additionally uncertain about whether to reimburse their loans. The risk of loan instalment can be adequately diminished by examining client behaviour ahead of time [4]. To assess whether another candidate can get a loan (an applied score), money related organizations commonly utilize a credit scoring model to examine and foresee the probability of a borrower defaulting on a current credit item (behavioral scoring) [5].

**Problem Statement:**

Credit scoring model predictions have now become an important component of the commercial world. In credit scoring, a customer’s creditworthiness describes their ability and willingness to repay a financial obligation (e.g. a loan). Credit scoring is a conventional decision model and its main focus is on risk approximation approach associated with credit products.

Currently, financial institutions are adopting various risk assessment tools and techniques for credit scoring systems to minimize the risk up to some extent [6]. The use of statistical tools for analyzing customer credit data helps to find out the default customer [7].

The number of compensations of credit scoring model incorporates such as establishing & fading credit risk and its concert is responsible for the profitability of the foundation [8, 9]. Various studies in the past have shown that ensemble-based approaches are effective in credit score assessment.

A major problem with these applications is the need to improve the scoring accuracy of loan selection. Improving the scoring accuracy of the credit decision by as much as a fraction of a percent can result in significant future savings and reduce the credit risk [10]. To accomplish even minor changes in loan scoring accuracy, the expert must look into non-standard classification architecture, as well as non-parametric statistical designs and classification trees.

**Aim and Scope of this project:**

The motivation behind this task is in the love of analysis and research made me to stick on this topic. This paper based on two research aims, first wants to compare the prediction of credit scoring system using machine learning and deep learning algorithms with knowledge distillation boosting technique for the financial industry. Second one is utilize on this credit scoring model to build a credit scorecard to assign credit scores for the existing customer and new customer. It will include exploratory data analysis, feature filtering, Weight of Evidence, Information Value, Binning, model building, and scoring.

Additionally, Knowledge distillation, aim to improving the predictive power of machine learning algorithms by extracting the knowledge contained in a complex model and injecting it into a more convenient model. When the cumbersome model has been prepared, we would then be able to utilize an alternate sort of training, which we call “Distillation” to exchange the knowledge from unwidely model to a little model that is progressively appropriate for deployment. An adaptation of this techniques has just been spearheaded by Rich Caruana and his colleagues [11].

The following are the areas that will be focused for this project:

1. Literature Review
2. Data collection & Preprocessing
3. Proposed Methodology and Model Implementation
4. Evaluation and Results
5. Building hybrid score card
6. Conclusion and Future Works

Scope of this project is to investigate the accuracy of the machine learning algorithms and Deep learning models via distillation for the credit scoring model and to benchmark their performance against the models currently under industry now. In this thesis we are using a datasets from Kaggle and UCI repository.

**Project Objectives:**

* To conduct the study about the Credit scoring models literature and highlight some of the techniques previously used and wide grow of the data presence, which led to surfacing of the statistical techniques.
* To focus on our methodology for the data preparation, missing data imputation, some feature selection techniques, and modelling techniques that will be apply on the data. We use the WoE binning and IV will be used for feature engineering and selection and are extensively used for credit risk modelling.
* To apply the models to our dataset and highlight all the results and discuss the accuracy of each model and to improve the accuracy we use the knowledge distillation technique.
* Evaluate the accuracy of the model using various metrics such as precision, recall, F1 score, and AUC-ROC score.
* Fine-tune the model by adjusting hyper parameters and repeating the evaluation until the desired level of accuracy is achieved.
* Deploy the credit score model in a production environment, where it can be used to make credit decisions in real-time.
* To provide recommendations for future study on hybrid credit score model and to present the overview of the key findings, and limitations.

**Expected Project Outcomes:**

* This project aims to develop a hybrid credit scoring model that can predict creditworthiness more effectively than traditional models and can be used to make credit decisions in real-time and improving the accuracy and reliability of credit decisions, and increase profitability for lenders, reduce the risk of default.
* The research will contribute to the development of accurate and efficient credit score detection models that can be applied in various industries.
* The study's findings will be valuable for businesses, researchers, and policymakers concerned with fraud detection and prevention.

**Limitations:**

* Internally developed credit scoring models often lack sophistication and usually have not been subjected to critical analysis of the statistical significance of the factors used to develop a credit score and a credit recommendation
* Professionally designed, tested and validated credit scoring models can be expensive, and can be hard to customize.
* Due to data observation limitations, the model will forecast incorrectly for a new occurrence in the test data set or in the industry because of the unknown recognized pattern.

By acknowledging these limitations, one can design and develop a more effective credit score card model for financial industry that takes into account these potential issues and works to address them for the future research.

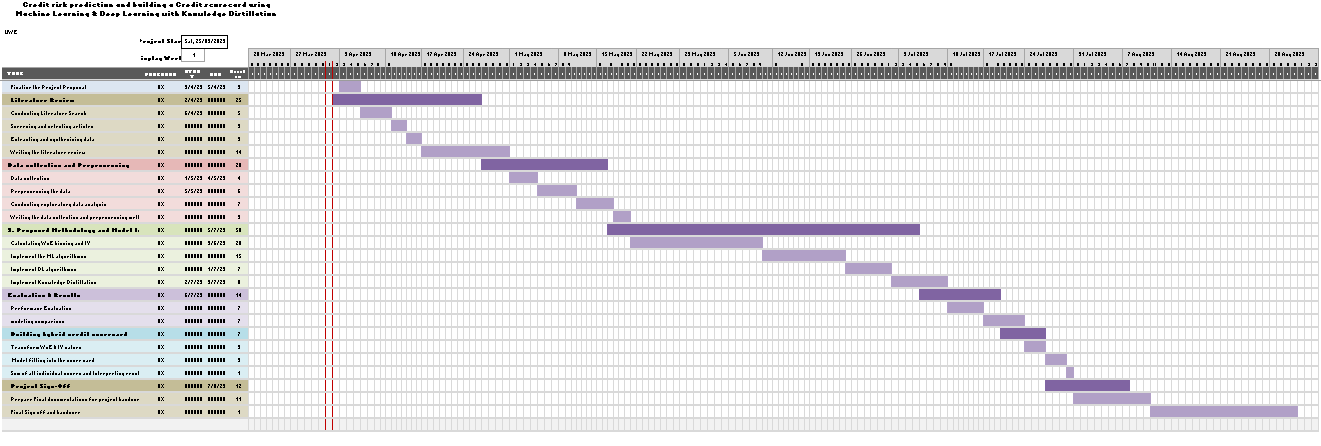
**Testing and Evaluation:**

* As per[12] the performance evaluation criteria such as Confusion Matrix (CM) or the Average Correct Classification (ACC) rate, the Estimated Misclassification Cost, Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), the Receiver Operating Characteristics (ROC) curve, GINI coefficient, and other criteria are all used in credit scoring applications under different fields.
* Commonly the mainstream of credit scoring applications either in accounting and finance or other fields have used the “average correct classification rate” as a performance evaluation measure, [13]. So model evaluation will be conducted that the average correct classification rate is an important criterion to be used, especially for new applications of credit scoring, because it highlights the accuracy of the predictions.

**Ethical Considerations:**

* The data sets used in this paper are from Kaggle's/UCI respository public data set, which is also a real data set after processing and is very popular in machine learning research and analysis [14].

**Project Plan:**



**References:**

[1] Šušteršič, M., Mramor, D., & Zupan, J. (2007). Consumer credit scoring models with limited data. Expert Systems with Applications, 36(3), 4736-4744.

[2] Bennouna, G., & Tkiouat, M. (2018). Fuzzy logic approach applied to credit scoring for microfinance in morocco. Procedia Computer Science,127, 274–283.

[3] Louzada, F., Moreira, F. F., & Oliveira, M. R. D. (2017). A zero-inflated non default rate regression model for credit scoring data. Communication in Statistics- Theory and Methods (1).

[4] S.Sampath and V. Kalaichelvi. (2014) “Assessment of mortgage application using fuzzy logic”. World academy of Science , Engineering and Technology, International Journal of Social, Behavioral, Education, Economic , usiness and Industrial Engineering. Vol 8, No 11.

[5] Djeundje, V. B., & Crook, J. (2018). Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards. European Journal of Operational Research.

[6] C. Serrano-Cinca and B. Guti´errez-Nieto, “The use of profit scoring as an alternative to credit scoring systems in peer-topeer (p2p) lending,“ Decision Support Systems, vol. 89, pp. 113–122, 2016.

[7] C. Liberati and F. Camillo, “Personal values and credit scoring: new insights in the financial prediction,“ Journal of the Operational Research Society, vol. 69, no. 12, pp. 1–21, 2018.

[8] Mester, L. J. (1997). “What’s the point of credit scoring?“ Business review, 3(Sep/Oct), 3-16.

[9] Thomas, L., Crook, J., & Edelman, D. (2017). “Credit scoring and its applications“. Society for industrial and Applied Mathematics.

[10] West, D. (2000). Neural network credit scoring models. Computers & Operations Research, 27, 1131–1152. 5, 7, 105, 113, 126

[11] Distilling the Knowledge in a Neural Network, Geoffrey Hinton, Oriol Vinyals, Jeff Dean, 2015

[12] [Performance Evaluation Criteria of Credit Scoring Models for Commercial Lenders-IJMTT, July 2019, akker Daniel K, Odundo F, Nyakinda J, Paul Samuel F ],

[13] West, D., Dellana, S., Qian, J. Neural network ensemble strategies for financial decision applications. Computers & Operations Research 32 (10): 2543-2559., 2010

[14] Credit Fraud Detection Based on Hybrid Credit Scoring Model -ICCIDS 2019, Keqin Chen,Amit Yadav, Asif Khan and Kun zhu