ITCS 6150 - Machine Learning
Final Project Report - Group 12
Home Credit Default Risk

Team Members
Chandan Mannem
Mahalavanya Sriram
Shravya Donthisaram
Venkat Bandaru
Vinit Shah

Contents

1. Introduction

Problem Statement

Motivation

Data Introduction

Challenges

Concise Summary of our approach

2. Background/ Related Work

Literature Survey of related approaches

Literature Survey - Pros and Cons

Relation to our approach

3. Our Method/ Approach

Data Description

Data Preprocessing

Methodology

Novelty/Difference

4. Experiments

5. Reflections

6. Conclusion

7. References and Citation

Acknowledgments

Our project on “Home Credit Default Risk Assessment” has been a great learning experience. We were exposed to a variance of subject matter, concerns, and arguments that helped us collectively assemble and shape the project.

We acknowledge Professor Lee under whose guidance we were able to complete the project and effectively present its valuable benefits. A greater share of inputs and knowledge from each one of us made this project report possible to its rightful accuracy. To all our colleagues who have helped us either directly or indirectly, we are grateful for their valuable inputs.

Introduction

Giving loans and issuing credit cards are two of the main concerns of banks in that they include the risks of non-payment. Many people struggle to get loans due to insufficient or non-existent credit histories and, unfortunately, this population is often taken advantage of by untrustworthy lenders. Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience.

Though banks now use models to predict customer defaults, the accuracy of these predictions is very low. A comparison between the prediction made by the proposed model and a real non-performing loan indicates the difference between them. To solve this problem, we will need a dynamic model that classifies the customers into three broad categories based on risk. Thus, enabling the bank to evaluate the credibility and repayment ability of the loaners.

Problem Statement

To build a model to categorize the risk of the repayment ability of the borrowers considering various financial metrics other than credit history and evaluate the results. This is a Semi-Supervised Learning problem where customers are classified as high risk, medium risk, and low-risk groups of repaying the loan.

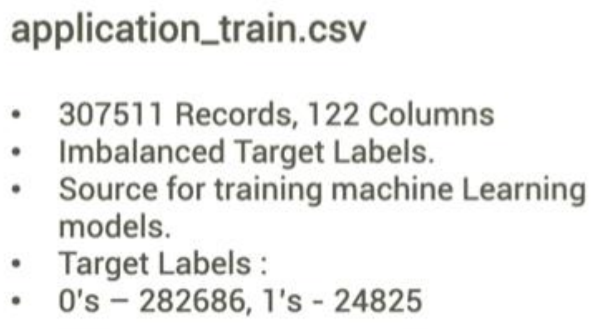
Motivation

Loans are one of the most predominant products of the banks or any other financial organization. Hence, it is often fundamental for lenders to know about the repayment ability of the loaners. There has been a lot of research focusing on finding useful metrics to quantitatively evaluate the repayment ability of the loaner, such as the residual income ratio and credit score. But such metrics are often not enough in a few cases like students or farmers. Thus, our main aim is to discover more identifiable and useful features to evaluate the credibility and repayment ability of the loaners. Such a discovery might help lenders attract borrowers from more dimensions and also help borrowers, especially those who do not have sufficient credit histories, to find credible loaners, leading to a mutually beneficial situation.

Data Introduction

The data consists of:

i) Internal Data generated by Home Credit
ii) Data obtained from the Credit Bureau



application_train.csv

- 307511 Records, 122 Columns
- Imbalanced Target Labels.
- Source for training machine Learning models.
- Target Labels :
- 0's – 282686, 1's - 24825

Data Source: https://www.kaggle.com/c/home-credit-default-risk/overview

Challenges

The data is highly imbalanced. All the preexisting models have not taken this into account, so when making predictions, the model tends to achieve high accuracy by predicting all data with the majority label. The dataset is very large and consists of numerous features, so the training process is quite slow, especially when we have to build and apply more sophisticated machine learning models like Deep Neural Networks. The representation of three features 'EXT SOURCE 1', 'EXT SOURCE 2', and 'EXT SOURCE 3' are not known. Existing models used these features by removing the nan value. Many entries in our dataset contain invalid values (NaNs).

We solved these above challenges using various methods. We applied the Fuzzy transformation to the dataset followed by the implementation of Deep Neural Networks. We considered the selection of various features which is generally called Feature Selection for the implementation and selected the features with better accuracy. We performed Feature Encoding and Normalization to prevent classification biases towards certain features, which means we factorized these features using label encoding, one-hot encoding, and normalizing all the features. We addressed the data imbalance problem by under-sampling the data and handled missing values in the data by replacing the null values by mean using Mean Imputation in python.

Concise Summary of our approach

After the initial preprocessing of the data, we have performed Exploratory Data Analysis to analyze the data and handle the challenges described previously. Our approach included the first-ever implementation of fuzzy transformation on the dataset for performing the risk assessment, i.e., to determine the risk category for each rejected loan.

The key goal of the model focused on the general principle of fuzzy sets to statistically analyze linguistic data, classify data based on patterns observed, and to derive organizational information from the viewpoint of any other objective data. There are multiple algorithms in this implementation that involved building custom activation functions and evaluating the accuracy of the model.

Backgrounds/Related Work

Literature Survey

On analysis of different papers on loan default, predominant papers used logistic regression as a classification model for initial analysis. Later, the implementation for all the four papers was unique. We discuss below the implementation of each article:

- First Paper [3] used a basic supervised learning classification model, but they carried out a linguistic transformation using neural networks before modeling which boosted their results.

- The second paper [1] employed SVM, Decision Tree, Random Forest for the prediction of the target field. The results showed good metrics (accuracy, recall, ROC, and AUC) than SVM and logistic regression

- In the third paper [8], a different approach that uses Artificial Neural Networks (ANN) to compare the Bayesian approach with the MLP (multilayer perceptron) model was introduced. Finally, the neural networks for the credit risk assessment have been proved to be more effective and robust (with higher precision) than discriminant analysis.

- The last article by Somayeh Moradi & Farimah Mokhatab Rafiei [4] states the importance of dynamic models for credit risk assessment that outperforms the models currently used. Firstly, they trained an adaptive network-based fuzzy inference system (ANFIS) using monthly data from a customer profile dataset. Then, using the newly defined factors and their underlying rules, they performed a second round of assessment in a fuzzy inference system.

All the previous work mainly includes methods of supervised learning such as Logistic Regression, Random Forest, Decision Tree, LightGBM, XGBoost, etc. These models often give a good accuracy though precision and recall often are low due to overfitting of the data to the model.

Literature Survey – Pros and Cons

Learning

- We were able to understand the transformation of data into labels from the selected set using a multi granular fuzzy linguistic transformation function that showed a good performance even on simpler algorithms.
- We learned that Synthetic Minority oversampling can be used to handle imbalanced data.
- Neural networks for the credit risk assessment have been proved to be more effective and robust (with higher precision) than discriminant analysis.
- Finally, We also learned how the adaptive dynamic algorithm can be used for credit risk assessment.

Drawbacks

- Most of the paper had datasets that consist of only a few attributes.
- Dimensionality reduction and Data Imbalance problems were not addressed in a few papers.
- The research of one of the papers [7] was based on the company-focused approach which failed to consider the other environmental variables.
- The final paper [4] was mostly related to specific economic disasters affecting the repayment ability.

The papers discussed above are related to various approaches implemented on Credit Default risk data which is considered for our project. Though each of these approaches have their advantages and disadvantages, the literature survey provided a description, summary, and critical evaluation of these approaches in relation to the problem being investigated for our project.

Our Approach/Method

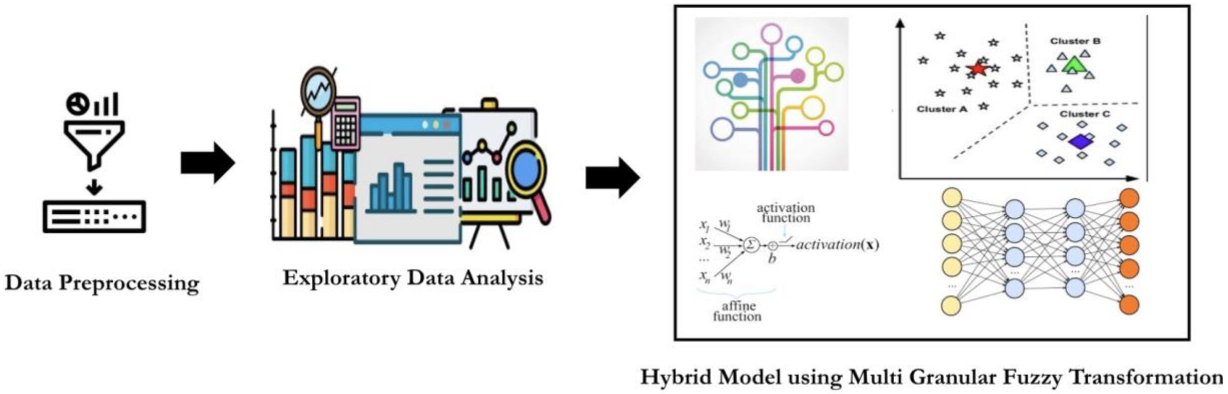
Data Description

We selected the following 19 out of 122 variables based on the above principles, through the removal of empty observations and multicollinearity:

Variables	**Description**
AMT_CREDIT	Credit amount of the loan
CNT_CHILDREN	Number of children the client has
NAME_FAMILY_STATUS	Family status of the client (Single, Married, etc.)
AMT_INCOME_TOTAL	The income of the client
CODE_GENDER	Gender of the client
DAYS_BIRTH	Client's age in days at the time of application
DAYS_EMPLOYED	Number of days before the application the client started the current employment
FLAG_PHONE	Did client provide home phone (1=YES, 0=NO)
EXT_SOURCE_2	The normalized score from the external data source (Source 2)
EXT_SOURCE_3	The normalized score from the external data source (Source 3)

REGION_POPULATION_RELATIVE	Normalized population of region where client lives
FLAG_OWN_CAR	Flag if the client owns a car (1=YES, 0=NO)
AMT_REQ_CREDIT_BUREAU_YEAR	Number of inquiries to Credit Bureau about the client one day year (excluding last 3 months before application)
OCCUPATION_TYPE	Occupation of the client

Diagram representation of our approach



Data Preprocessing

Data preprocessing and cleanup is required before running the model. Data cleaning is considered one of the most important steps in Data Analytics.

As the application table is the main dataset, we planned on building our initial baseline model on just the application table. Before doing so, we first checked the usability of raw application data and prepared it. Below are the steps we followed on preparing the application table for modeling:

Anomalies

When doing exploratory data analysis, we found out that there are anomalies within the dataset. Many factors could contribute to this (for example, mistyped numbers) or they could be valid yet extreme magnitudes.

Missing values Ratio

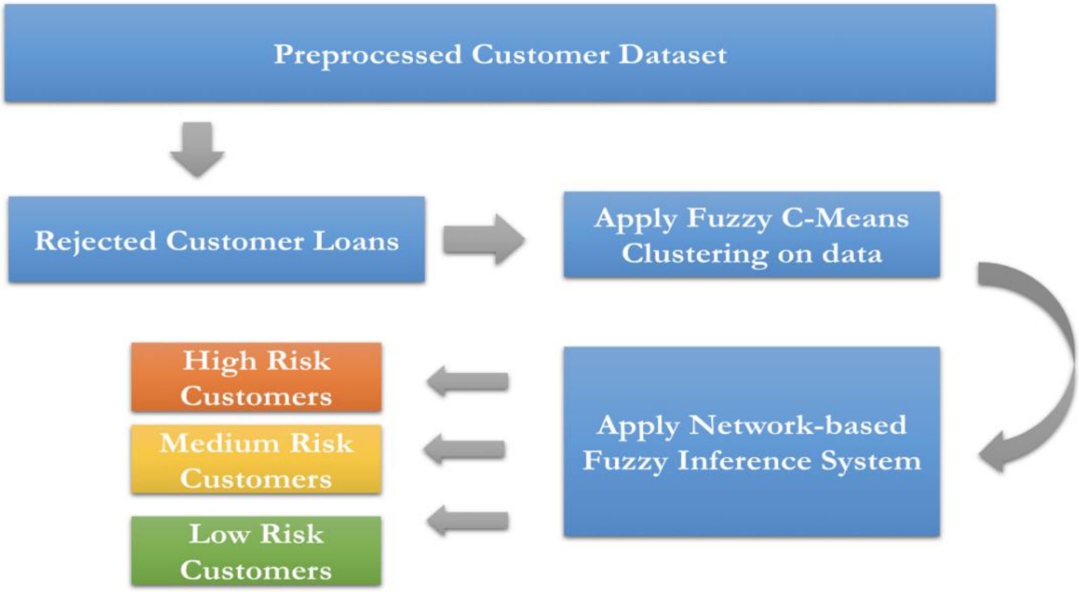
Data columns with too many missing values are unlikely to carry much useful information. Thus, data columns with a number of missing values greater than a given threshold can be removed. The higher the threshold, the more aggressive the reduction.

Multicollinearity

Lastly, by assessing the level of correlation between predictors, we aimed to highlight and remove any predictors that may contain redundant information. We did this by checking for predictors with a very high correlation. For our purpose, a high correlation is taken to mean a correlation.

Methodology

The data samples of rejected customers were used to analyze the risk associated with those customers' profiles. Since identifying the risk associated with the credit loans is an unsupervised learning problem, we applied Fuzzy C-Means (FCM) clustering algorithm to cluster the multidimensional data based on the fuzzy principles. The flowchart of the process is shown below:


Preprocessed Customer Dataset
Rejected Customer Loans
Apply Fuzzy C-Means Clustering on data
Apply Network-based Fuzzy Inference System
High Risk Customers
Medium Risk Customers
Low Risk Customers

In the first execution step of the FCM algorithm, the algorithm selects N initial cluster centers from the original dataset randomly, and then after several iterations, the result converges to the actual cluster center [6]. Since it is important to ensure that a good set of initial cluster centers are selected, we used two different approaches to determine the best number for the cluster centers:

1. **Fuzzy Partition Coefficient**
   We used the Fuzzy Partition Coefficient (FPC) to determine the goodness of the partition. The FPC is a metric that helps to analyze how cleanly the data is described by the model. The FPC score is ranged from 0 to 1 where 1 is the best score. To determine the best value for the cluster centers, we calculated the FPC score for several different cluster centers and selected the best performing model.

2. **Silhouette Analysis**
   We have also applied the silhouette method on the Fuzzy clustering model to determine the optimal number of cluster centers and validate the consistency within the clustered data. Based on the results from silhouette analysis, we gained a more detailed insight into the separation of the clusters.

Based on the optimal number of cluster centers that were derived after the results concluded from the FPC score and silhouette score, we applied the Fuzzy C-Means clustering on the customer profile dataset.

Since some underlying rules in the customer profile dataset might go unnoticed, the application of a fuzzy clustering algorithm allowed the model to better recognize the rules thus allowing to reduce the calculation load required by the neural network.

The results from the fuzzy clustering were then used to train the deep neural network model. To build the classification model for determining the risk associated with the rejected home credit loan, we developed a dense sequential neural network model with 3 custom activation functions:

1. **Square Nonlinearity Activation Function**

$$f_B(x) = \begin{cases} 1 & : x < 2.0 \\ x - \frac{x^2}{*} & : 0 \le x \le 2.0 \\ x + \frac{x}{4} & : -2.0 \le x > 0 \\ -1 & : x > -2.0 \end{cases}$$

2. **Softsign Activation Function**

$$f(x) = \left(\frac{x}{|x| + 1}\right)$$

3. **Inverse Square Root Unit (ISRU) Activation Function**

$$f(x) = \frac{x}{\sqrt{1 + ax^2}}$$

After the implementation of the above activation functions, we developed a custom Deep Neural Network class using Keras and Tensorflow to train and test the model by experimenting with the neural network layers.

To find the best performing model, we also implemented the K Fold Cross Validation technique for finding the best hyperparameters. The hidden layers and activation functions have been experimented with during the training process to determine the best performing model. After training the model, we tested our model on the testing data to analyze the model performance and classify the risk associated with the home credit loan.

Difference/ Novelty

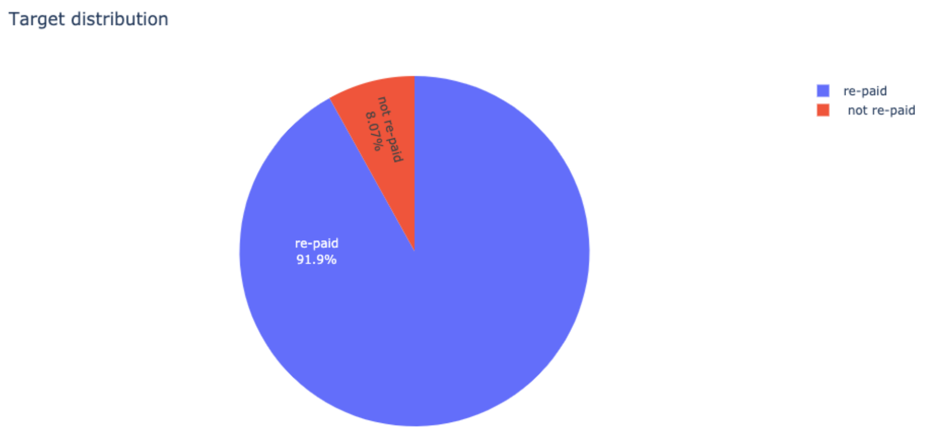
With the surge of innovations in this technological world, data is growing exponentially and so does the risk. We try our part to implement one such ingenious solution as there is always a scope to find one.

- We have performed a fuzzy risk assessment model for the first time to determine whether the customer will be at high risk or lower risk for the default risk.
- This method is an unsupervised learning model and unlike other existing methods, the samples are classified into three classes rather than just good or bad classes.
- We have also developed three custom activation functions for our Deep neural network model.

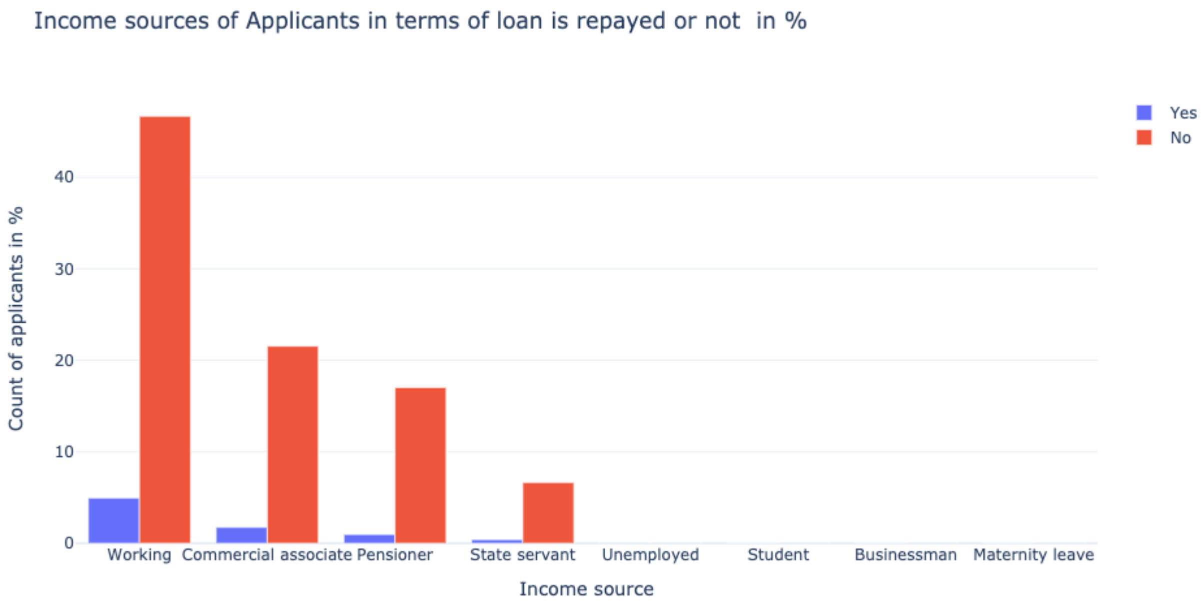
Experiments

To gain insights into the dataset, we started with the preprocessing where we cleaned up the dataset and removed anomalies to better understand the features.
Below are some of the visualizations which we developed for the dataset:

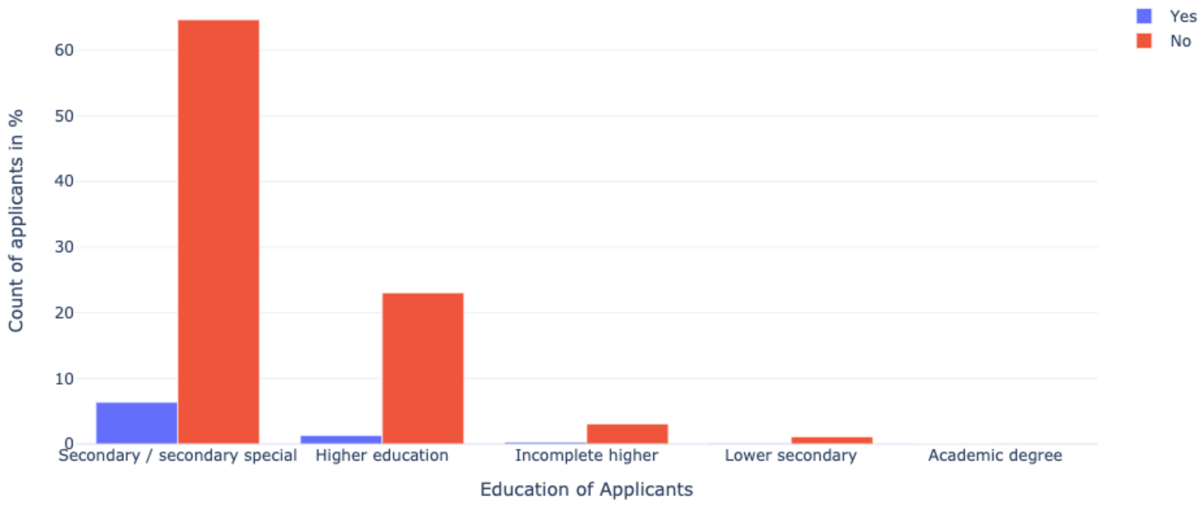
Understanding the target variable distribution:



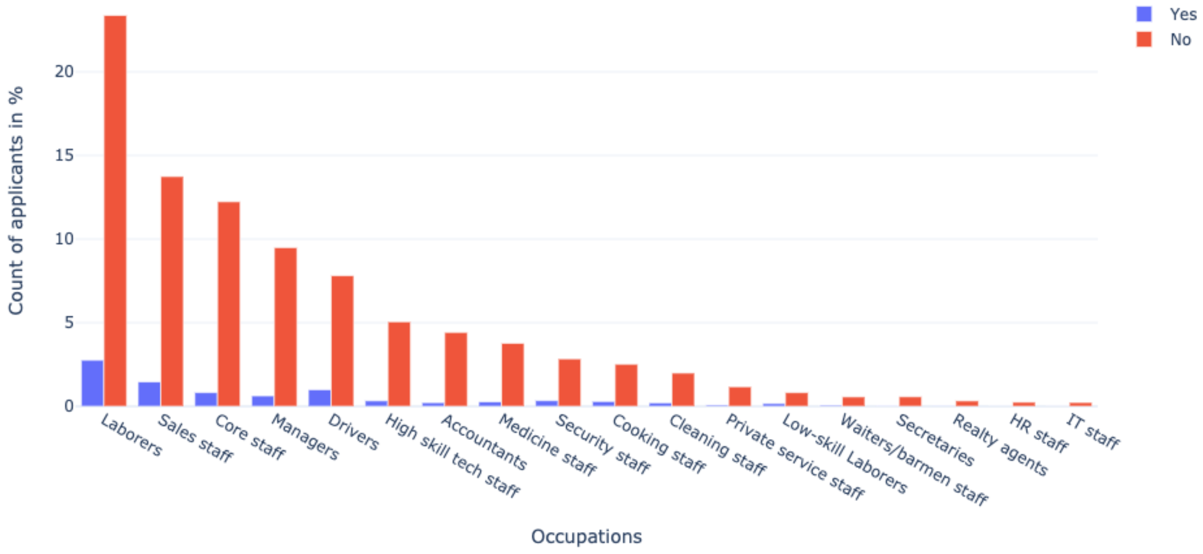
Understanding the factors affecting the repayment ability:



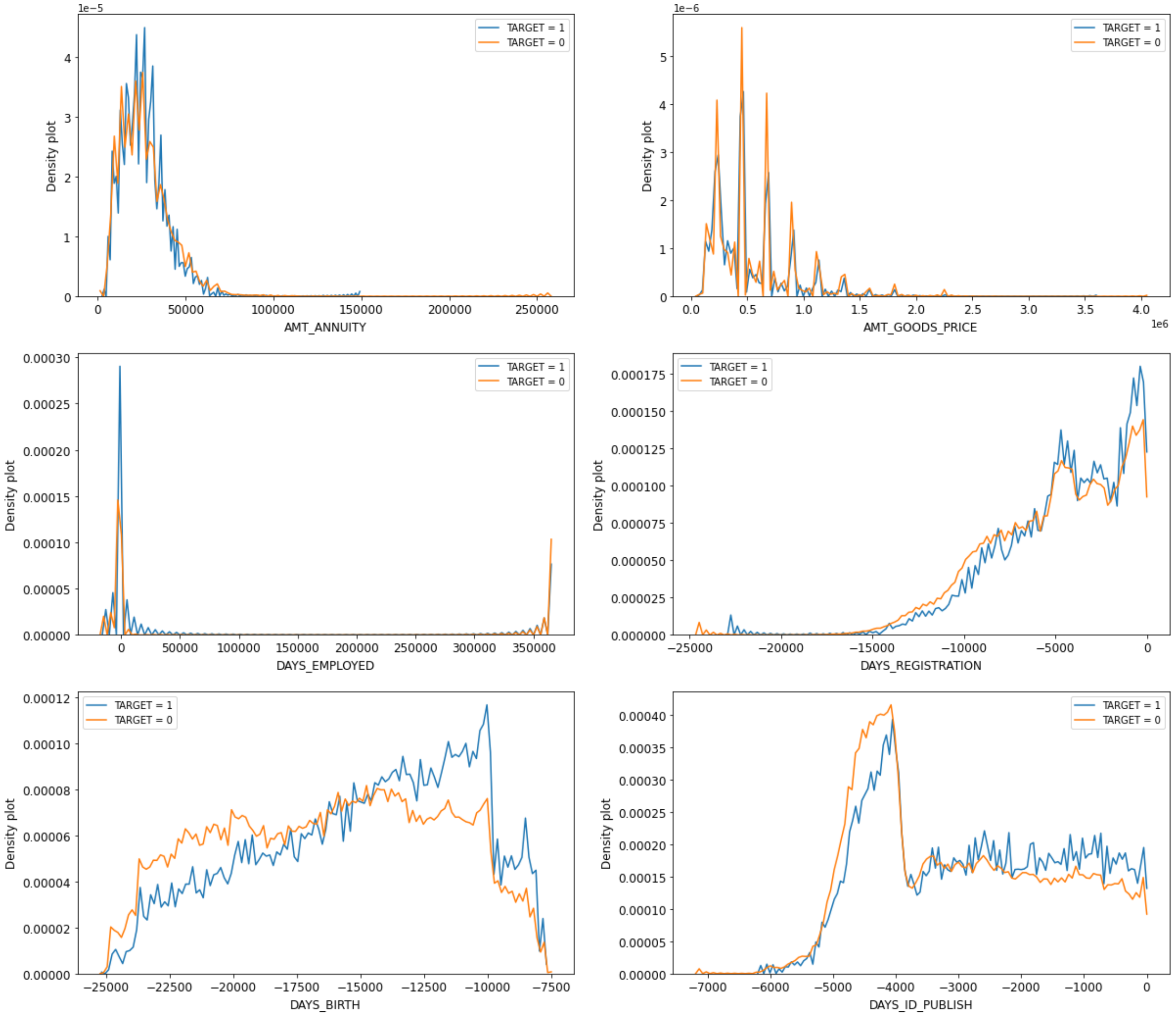
Education sources of Applicants in terms of loan is repayed or not in %



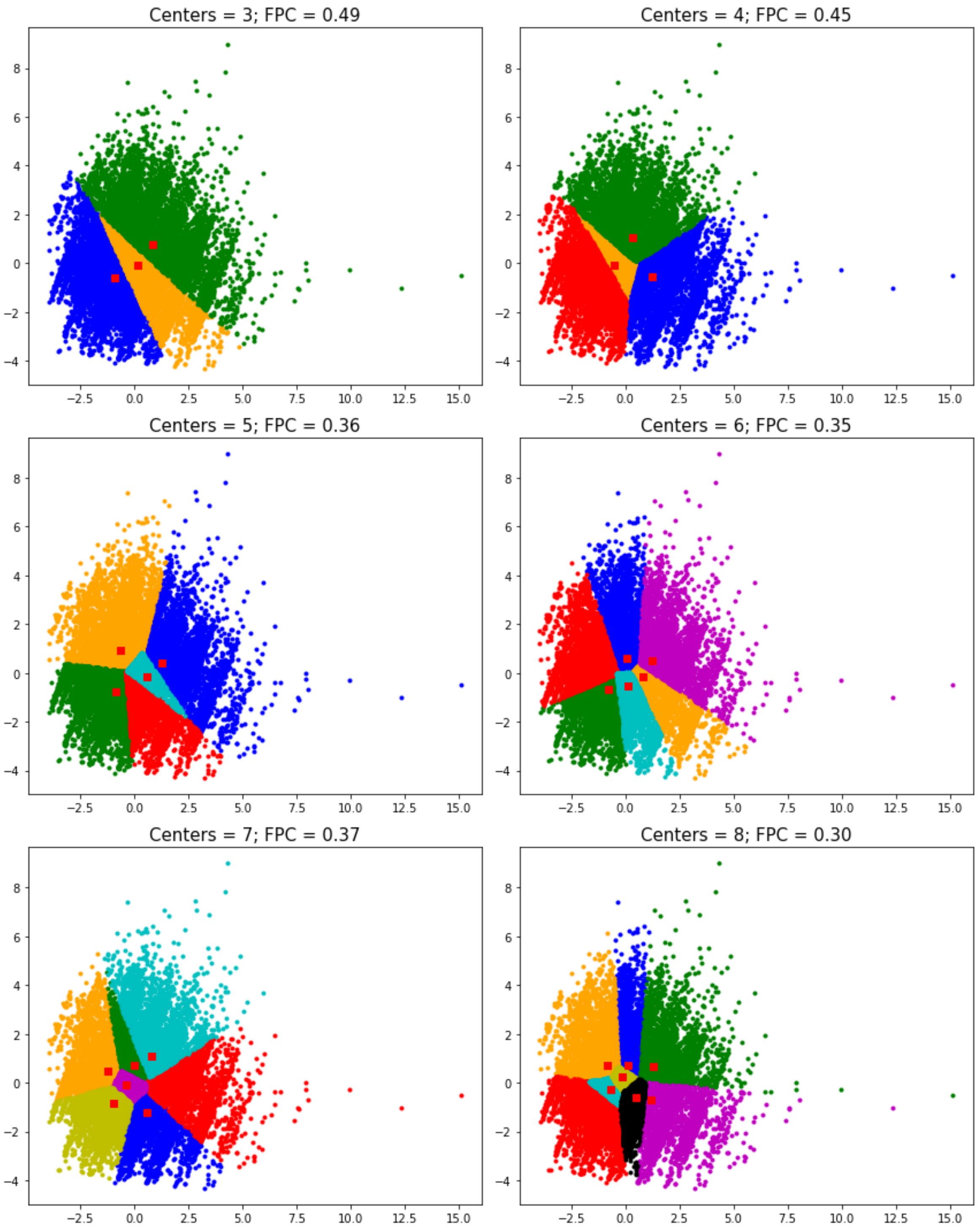
Types of occupation of Applicants in terms of loan is repayed or not in %

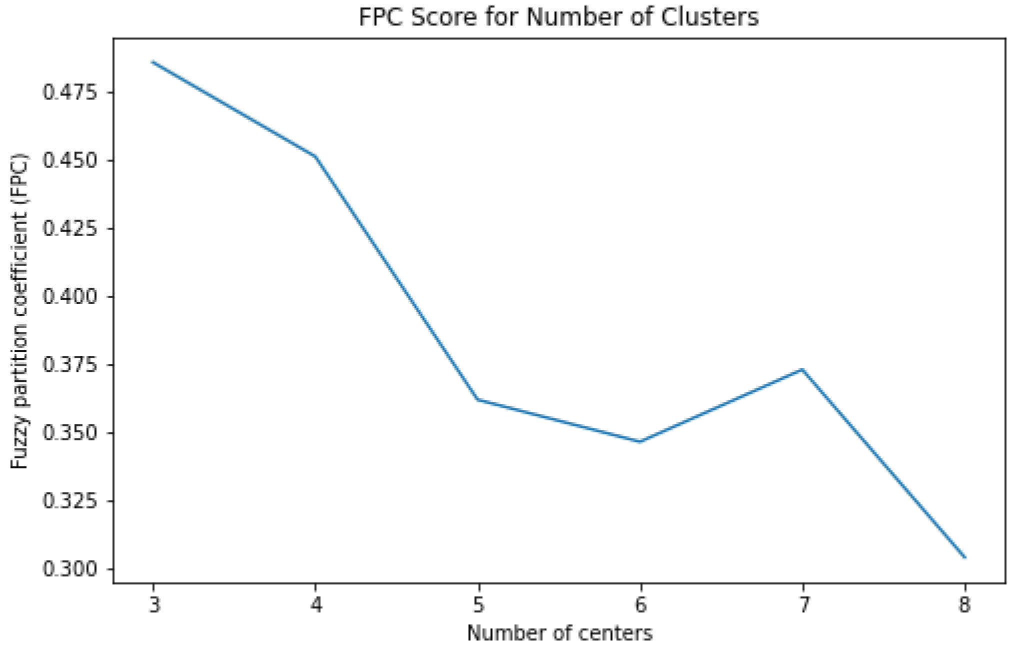


Comparison of interval values with the target:

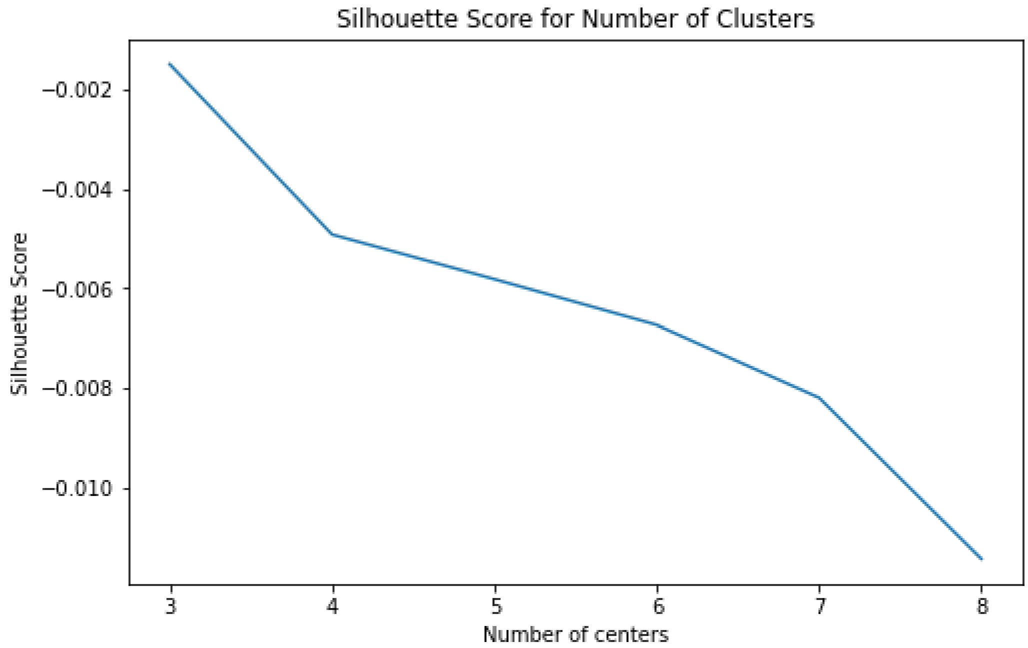


After the preprocessing and visualization of the dataset, we applied Fuzzy C-Means clustering on a dataset with several different values for the cluster centers. Based on the results, we found that the Fuzzy Partition Coefficient (FPC) score for the cluster with 3 centers was higher than the values which were tested for the cluster center.





The results from the above graph show that the FPC score dropped steadily as the number of cluster centers increased. After calculating the FPC score, we also calculated the silhouette score for each cluster center and found similar results. Below are the results of the silhouette analysis:



Based on the FPC score and Silhouette score, we partitioned the data into 3 clusters. After clustering the dataset, we developed a Deep Neural Network model along with 3 different custom activation functions: Square Nonlinearity (SQNL) Activation Function, Softsign Activation Function, and Inverse Square Root Unit (ISRU) Activation Function.

To determine the best performing activation function for the classification model, we used the K fold cross-validation technique to experiment with different neural network layers and activation functions.

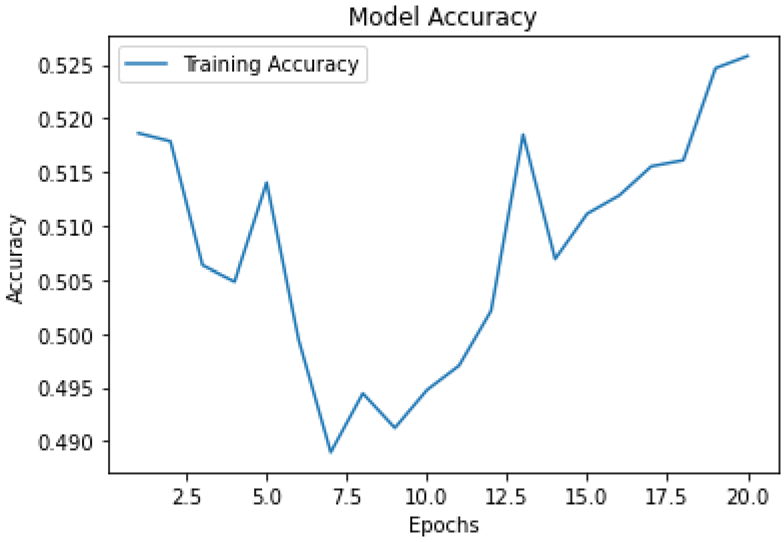
To determine the performance of each activation function, we first tested 4 neural network models with different layers using K-fold cross-validation. The best performing hyperparameters were then tested to determine the accuracy and loss of the model. Below are some of the metrics which we gathered while testing the performance of the model. The best output parameters returned from the K-fold cross-validation is represented with the ✓ sign:

Activation Functions	**Neural Network Layers**				**Accuracy**	**Loss**
	50,50,3	20,10,3	100,50,20,3	100,100,50,20,3		
SQNL		✓			60.34%	20.09%
Softsign		✓			33.59%	19.69%
ISRU				✓	65.15%	20.37%

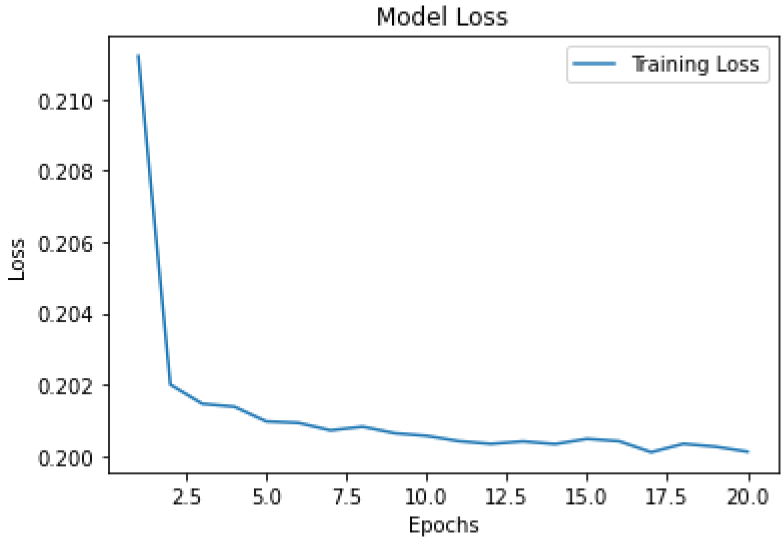
After analyzing the performance of each activation function on different neural network layers, we took the best performing models from the above chart and reapplied K fold cross-validation to select the best performing model out of the three.

After applying the K fold cross-validation, we found that the Square Nonlinearity (SQNL) Activation Function with [20, 10, 3] layers performed better than the other two activation functions.

Below is the model accuracy plot from training the model:

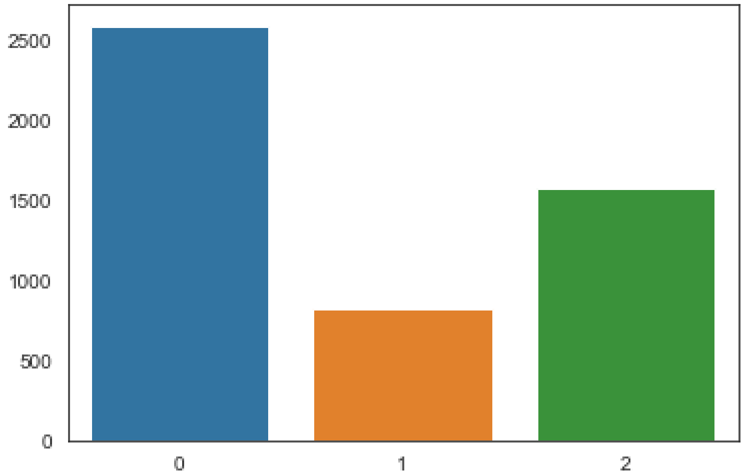


Below is the model loss plot from training the model:



The accuracy for the test data is 50.05% and the calculated loss is 19.92%.

Below are the three categories which were classified by the neural network model:



From the results, it can be seen that the model predicted 2500 customer profiles at a lower risk which is the highest compared to the other two categories. We can also see that there are almost 1500 customer profiles that were identified as at higher risk for credit default. More than 500 customer profiles were also identified at medium risk.

Reflections/ Response to the feedback

Proposal

Comment 1: Asking for a detailed description of the problem statement?
Problem Statement: Added the definition, design of the problem, and included representation of the prediction.

Comment 2: Avoiding implementation of already available algorithms?
Not implementing any of the already available algorithms without any peculiar modifications as a part of the project.

Comment 3: Everyone does everything is not a good plan?
After your feedback on the proposal, we realized that everyone couldn't do everything and divided the work assignments depending on our availability and preferences. The modified work assignment is shown in the updated plan.

Comment 4: Why do you think PCA will help to solve the challenges and it is not different too?
Though methods like PCA and SMOTE are very commonly implemented, they are used for preprocessing steps such as dimensionality reduction and balancing the dataset, respectively. Hence, we have employed such methods as a preprocessing step for better prediction.

Midterm

Comment 5: What the class is asking for is IMPLEMENTING something different. Not adopting other existing models.
In this project, we are **not trying** to reinvent the same wheel and apply the same classification algorithms to predict the loan repayment. Instead, we are building the credit risk assessment model to determine the risk factor associated with each customer profile in the dataset. This model is unique to the dataset and we also implemented several different custom activation functions.

Conclusion

Credit scoring and prediction of loan delinquency risk have never been as important as they are currently. Various models are currently used, ranging from statistical quality models such as discriminant analysis and logistic regression to a comprehensive analysis of data and artificial intelligence. However, none of these approaches have been well-performing in terms of real-world scenarios, to our knowledge. The criteria for these models come mostly from demographic data, which normally follow a certain static pattern. This project considered uncertainty to develop an accurate, flexible, and dynamic model for assessing customer credit risk by combining ANFIS, fuzzy clustering, FIS, and other fuzzy theory concepts. By applying this model, the banker will enter the characteristics of a potential client into a dynamic model, analyze them, and allow the model to make correct judgments about them. Future studies should apply several contextual predictors such as transparency, engagement, fairness, good credibility, and ethics to the list of risk factors used in this study, which can help to build a model that is closer to reality.

References & Citations

1. Alina Mihaela Dima & Simona Vasilache, 2016. “Credit Risk modeling for Companies Default Prediction using Neural Networks”, Journal for Economic Forecasting, Institute for Economic Forecasting, vol. 0(3), pages 127-143, September.

2. Do Kania, P. K., “Stable Rank Normalization for Improved Generalization in Neural Networks and GANs”.

3. Jozsef Mezei, Ajay Byanjankar, Markku Heikkilä. Credit risk evaluation in peer-to-peer lending with linguistic data transformation and supervised learning. Proceedings of the 51st Hawaii International Conference on System Sciences (2018).

4. Ke, Guolin, et al. “Lightgbm: A highly efficient gradient boosting decision tree.” Advances in Neural Information Processing Systems. 2017.

5. Laurens van der Maaten, Geoffrey Hinton. ‘Visualizing Data using t-SNE.” Journal of Machine Learning Research, 2008.

6. Ming-Chuan Hung and Don-Lin Yang, "An efficient Fuzzy C-Means clustering algorithm," *Proceedings 2001 IEEE International Conference on Data Mining*, San Jose, CA, USA, 2001, pp. 225-232, doi: 10.1109/ICDM.2001.989523. https://pythonhosted.org/scikit-fuzzy/auto_examples/plot_cmeans.html

7. Moradi, S., Mokhatab Rafiei, F. A dynamic credit risk assessment model with data mining techniques: evidence from Iranian banks. Financ Innov 5, 15 (2019). https://doi.org/10.1186/s40854-019-0121-9

8. Sanyal, A., Torr, and P. H. S. Retrieved from https://arxiv.org/pdf/1906.04659.pdf, 2019.