

Applied Machine Learning Final Project: Home Credit Default Risk

Indiana University

December 2022

Group 11

Anuj Mahajan

Shubham Jambhale

Shashwati Diware

Siddhant Patil

Team Members:

Shubham Jambhale
sjambhal@iu.edu



Siddhant Patil
sidpatil@iu.edu



Anuj Mahajan
anujmaha@iu.edu



Shashwati Diware
sdiware@iu.edu



Contents

- Four P's
- Project Description
- Overview of modelling Pipelines
- Modelling Pipeline Flow
- Results and discussion (Accuracy, AUC, Kaggle)
- Conclusion

Four P's

- **Past:**

- The HCDR Project, which uses a variety of financial and nonfinancial variables to determine whether borrowers will fail or not.
- We performed EDA and feature engineering to develop baseline models (Logistic, Decision, Random Forest, etc) to enhance results.
- AUC and the Confusion Matrix were used to assess the accuracy.
- We tuned the hyperparameters of the models and found the best parameters using Grid Search.

- **Present:**

- We built the Multi Linear Perceptron using PyTorch.
- In this project, to visualize the results of training in real-time we used Tensor board.

Four P's

- **Planned:**

- In near Future we plan to perform and use different Activation Functions with various different combinations to attain more reliable and accurate results.
- Try to reduce and minimize the loss and error rate as low as possible.

- **Problems:**

- There was a time constrain to run more experiments and different combination to achieve our optimal solution.
- Some computational barriers restricted us to perform some experiments and run at a larger expense.

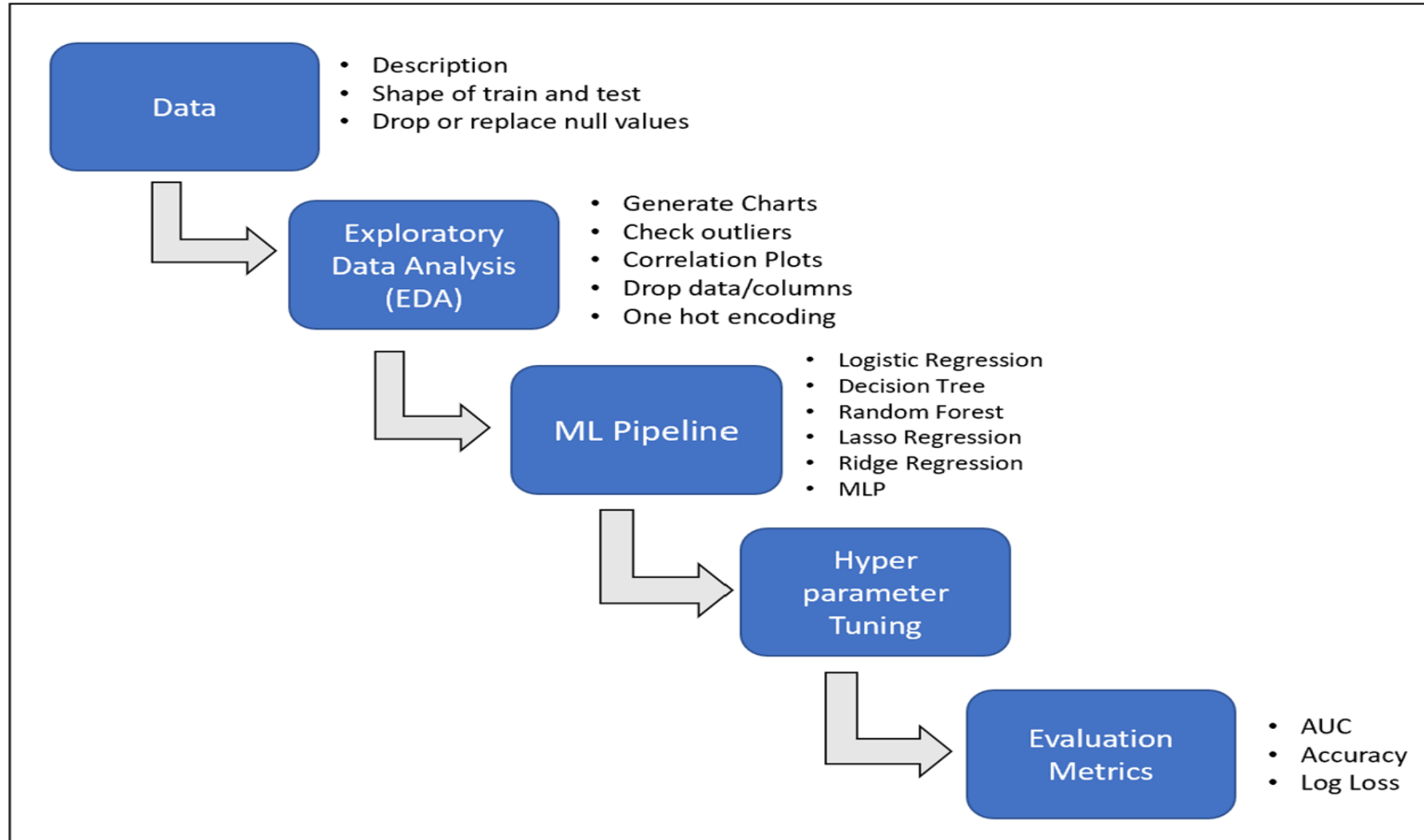
Project Description:

- The object of the Home Credit Default Risk (HCDR) project is to predict the repayment abilities of the financially under-served population.
 - The well-established prediction is necessary for both Home Credit and borrowers.
 - Lend money to whom can pay back and give them a chance to build credit.
- We trained and assessed a number of potential models before selecting the best one.
 - Our potential models include Random Forest, Decision Making Trees, Logistic Regression, Lasso Regression and Ridge Regression.
 - In this Phase we tried different MLP combinations to achieve the current results considering the accuracy and the AUC score.
 - Accuracy, AUC score, confusion Matrix and BCE loss are just a few measures we employ to evaluate the model precisely.

Modelling Pipeline:

- The goal is to predict whether the borrower is a defaulter or not. Phase 4 involves working on the Multilayer Perceptron Model with its combinations to achieve desirable results.
 - Preprocess the entire dataset.
 - Prepare the MLP model with all its parameters to achieve some output.
 - Process this Multilayer Perceptron model with different Activation Functions.
 - Using Activation Functions such as Relu and Sigmoid for the prediction.
 - At the end, analyze these models' using measures such as Accuracy, AUC, and loss functions.
 - Perform the above steps through multiple iterations.

Modelling Pipeline Flow:



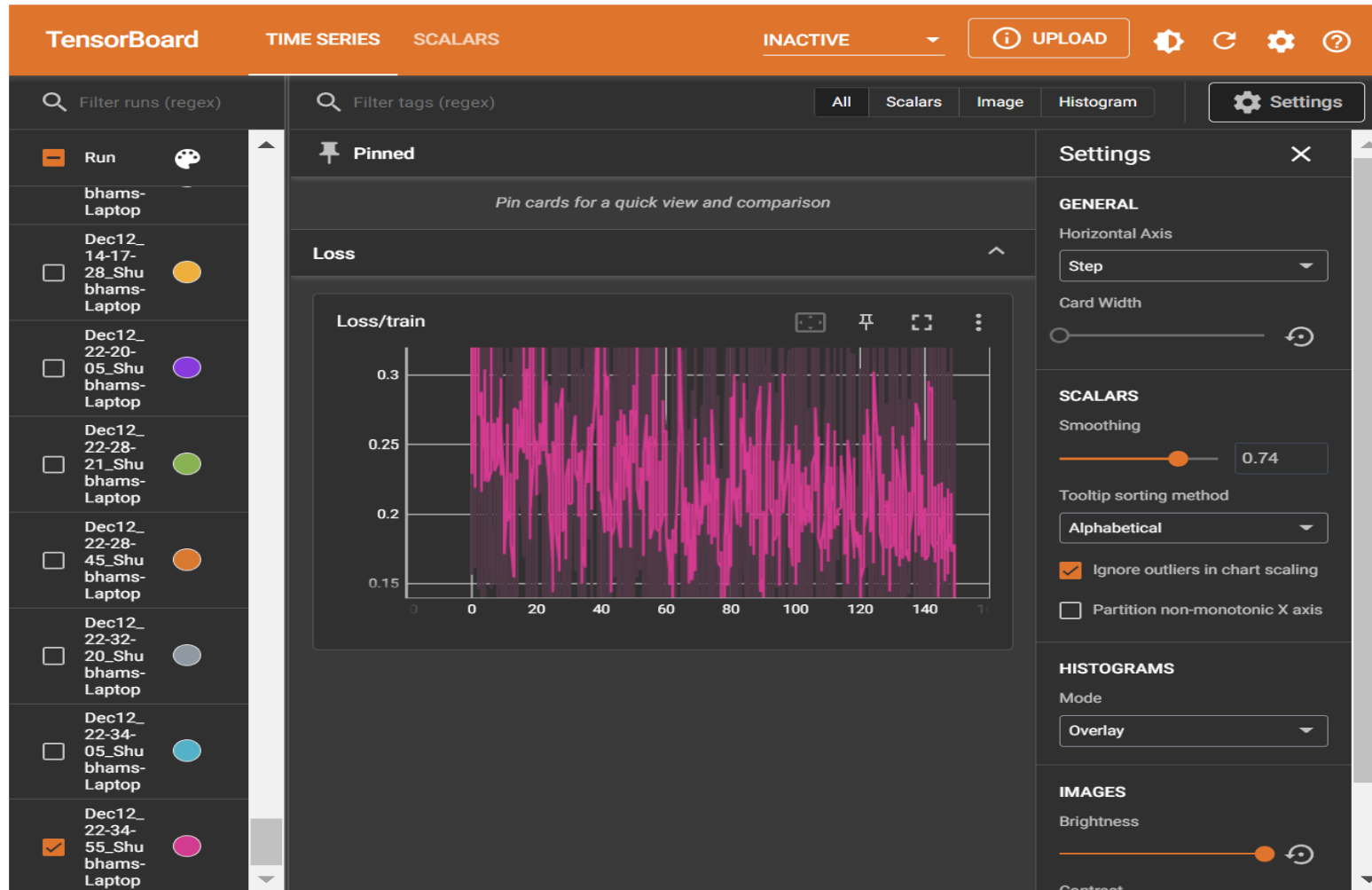
Result and Discussion:

- The overall accuracy of the decision tree grew significantly and reached 92%. Hyper tuned Decision Tree is the best-fit algorithm since it surpasses other models in Phase 3.
- In this Phase 4, we implemented deep learning, and our model was a multi-layer perceptron.
- We then generated a visualization of the loss function and accuracy using Tensor Board to visualize our training model.
- Our accuracy for the multi-Layer perceptron was out to be 92.4% which is quite efficient and overall good for the given data.
- We received Test AUC of 60 % for the Pytorch MLP

	ExpID	Train Time	Test Time	Accuracy	AUC	Comments
0	Multi Layer Perceptron	363.6610	0.0234	0.915039	0.601393	150 - ReLU + Sigmoid
1	Multi Layer Perceptron	366.6223	0.0326	0.932617	0.500000	150 - ReLU
2	Multi Layer Perceptron	160.7423	0.0156	0.920000	0.527774	100 - Sigmoid

	ExpID	Time	Accuracy	Valid Acc	AUC	Comment
0	MLP (Single hiddne layer)	15.8018	0.924275	0.909667	0.778271	MLP (Single hiddne layer)
1	MLP (multiple hidden layer)	15.8018	0.925275	0.909667	0.787373	MLP (Multiple hiddne layer)

TensorBoard:



Result and Discussion:

Below are our best submission from each phase 2 ,3, 4

Submissions

You selected 0 of 2 submissions to be evaluated for your final leaderboard score. Since you selected less than 2 submission, Kaggle auto-selected up to 2 submissions from among your public best-scoring unselected submissions for evaluation. The evaluated submission with the best Private Score is used for your final score.

0/2

■ Submissions evaluated for final score




All

Successful

Selected

Errors

Recent ▼

Submission and Description	Private Score ⓘ	Public Score ⓘ	Selected
 Submission4.csv Complete (after deadline) · now	0.50593	0.50461	<input type="checkbox"/>
 Submission3.csv Complete (after deadline) · 1s ago	0.71673	0.73086	<input type="checkbox"/>
 Submission2.csv Complete (after deadline) · 1m ago	0.50195	0.50271	<input type="checkbox"/>

Conclusion :

- During Phase 2 we concluded that the best model for this dataset will be logistic regression having the highest accuracy of 91.9 %.
- In Phase 3 we conclude that the decision tree model performs the best out of all the other hyper tuned models.
- As part of phase 4, we implemented Multi-Layer Perceptron with activation function ReLu and Sigmoid along with multiple hidden layer settings, we got the result that ReLu and Sigmoid together can give us the effective result.
- We have determined that the Multi-Layer Perceptron's accuracy for the provided dataset using PyTorch was 92.4%, which is extremely effective and generally good.
- With all the experiments taken into consideration Decision Tree and Lasso Regression turned out to be the best performing models. Additional experimentation on MLP model could have yielded better results.