AMERICAN EXPRESS CAMPUS SUPER BOWL

Team: THUNDER BUDDIES

- Siva Sankar S
- Keerthy Babu D
- Tejashaarav S

OUR TEAM

Siva Sankar S

- NLP engineer intern at Inscripta.AI
- AstraZeneca AI Hackathon Winner
- Inter IIT Tech-Meet: Contingent
- Upcoming Data Scientist intern at UPL, ltd.

Keerthy Babu D

- Upcoming analyst intern at American Express
- Data analyst intern at Stelllapps
- Project member at Computer Vision Club, CFI IITM

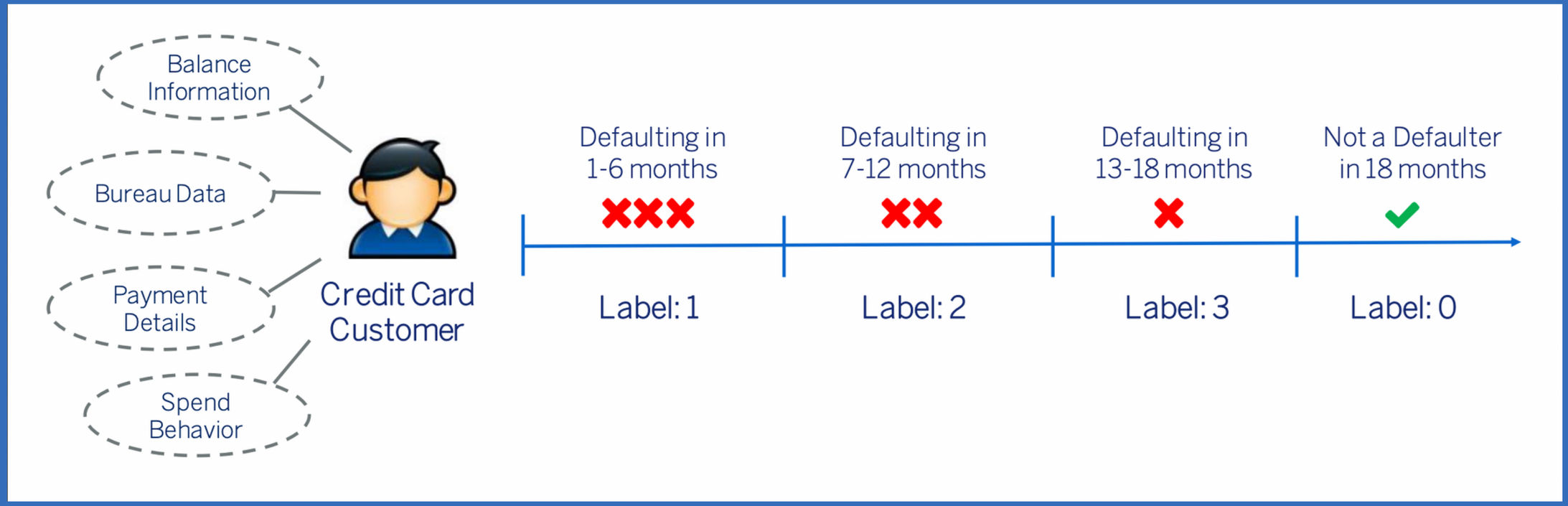
Tejashaarav S

- SDE, MLOps Intern at Softtech Engg, Pune
- Strategist at Team Sahaay, IITM
- Upcoming Analyst Intern at Dr.Reddy's Labarotaries

TEAM DETAILS

Name	College	Course	Batch Year	Roll No	Mobile Number	College Email ID
Siva Sankar S	IIT Madras	B.Tech in Chemical Engineering	2024	CH20B103	9283217898	ch20b103@smail.iitm.ac.in
Keerthy Babu D	IIT Madras	B.Tech in Chemical Engineering	2024	CH20B059	7483013164	ch20b059@smail.iitm.ac.in
Tejashaarav S	IIT Madras	B.Tech in Chemical Engineering	2024	CH20B107	9080783016	ch20b107@smail.iitm.ac.in

PROBLEM STATEMENT



- 188 features, 4 classes

Introduction

STAGES

The segments of our overall analysis are shown in brief here.

01 Data Pre-processing & Analysis

Exploring the data

Modelling 02

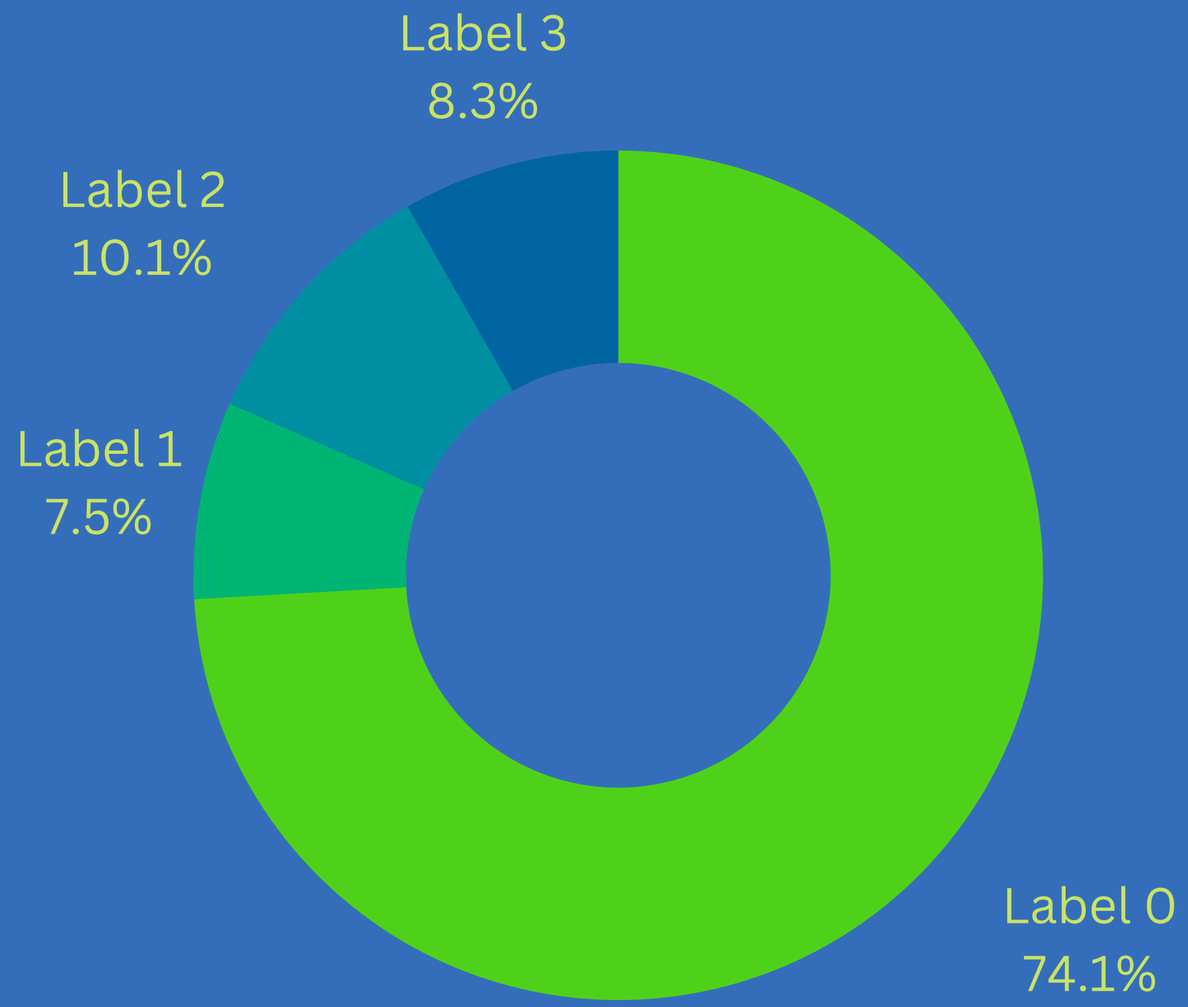
Things we tried out

03 Best submission

Final Submission

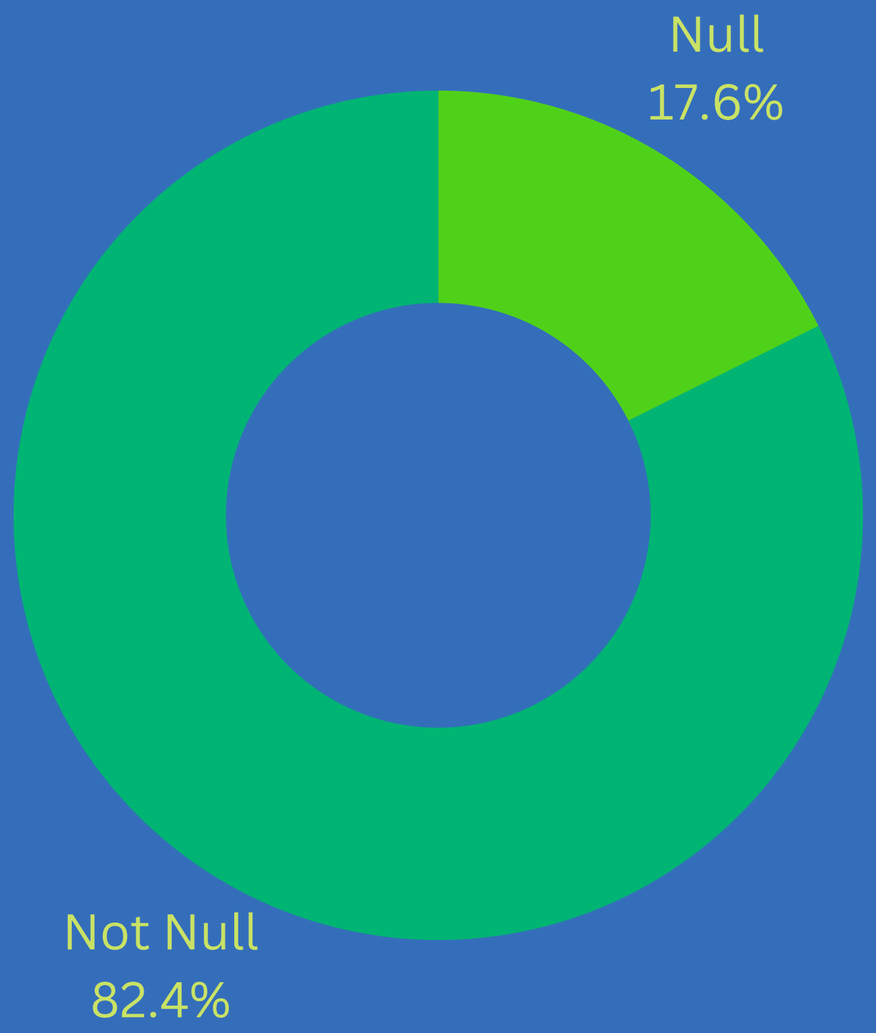
DATA IMBALANCE

- The data is highly skewed with label 0 being the majority class (74.11 %) and label 1 being minority class (7.52 %).
- We tried a lot of methods to handle this imbalance such as synthetic data point generation,(SMOTE), down-sampling, etc.
- And we found out that changing the original distribution improved F1 scores but overall accuracy is also reduced.

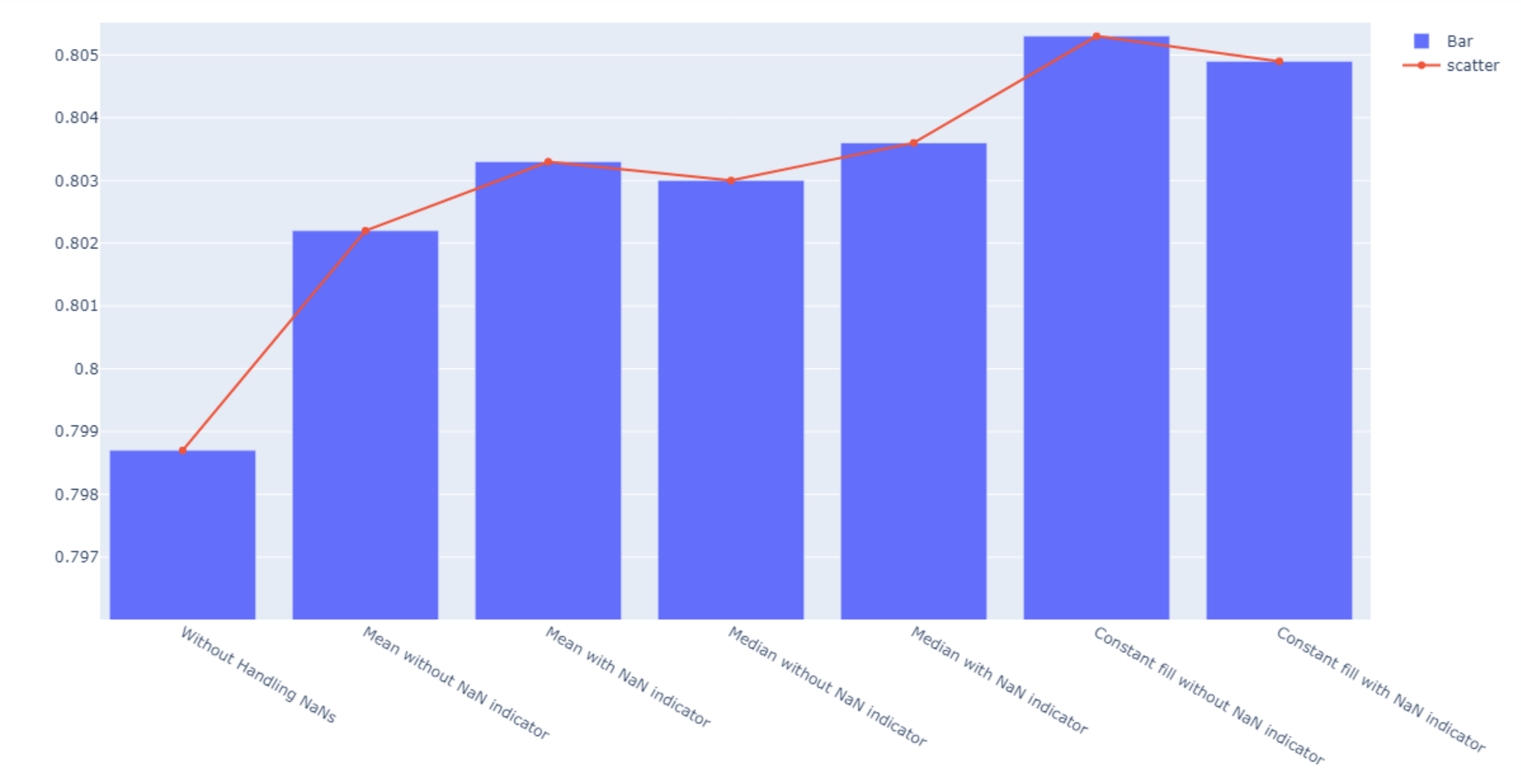


NAN VALUES

- 17.64 % of train data is null value.
- In order to handle the null values we conducted various experiments.
- We tried imputing null values with constant, mean, most frequent value, median and nearest neighbour.
- Our observation is that constant fill gives the optimal accuracy.



HANDLING NANS



FINAL SET OF PREPROCESSING

After experimenting and trying out a lot of preprocessing methods this is the flow we have finally chosen for preprocessing.

- Simple Imputation with constant value of 0.
- Quantile transformer to uniform distribution
- One hot encoding the categorical variables

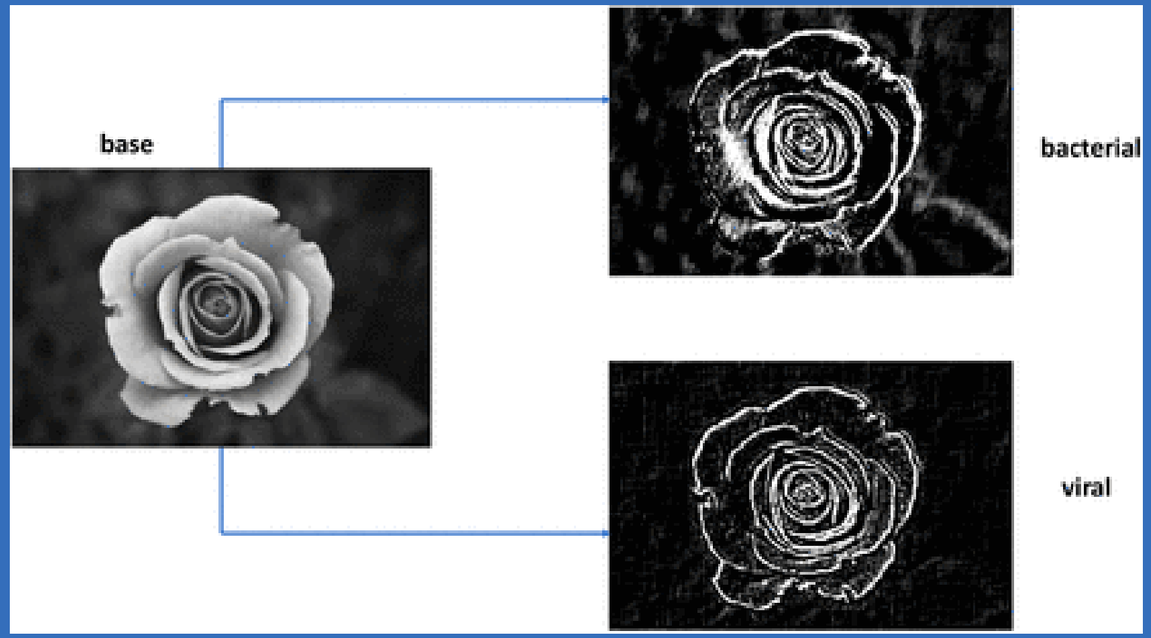
REFACTORING INTO COMPUTER VISION PROBLEM

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

A novel method for classification of tabular data using convolutional neural networks

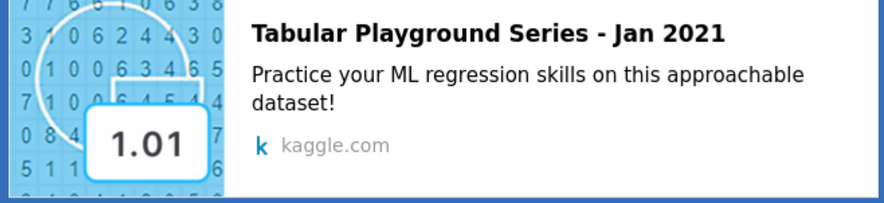
Convolutional neural networks (CNNs) represent a major breakthrough in image classification. However, there has not been similar progress in applying CNNs, or neural networks of any kind, to classification of...

bR bioRxiv / May 3, 2020

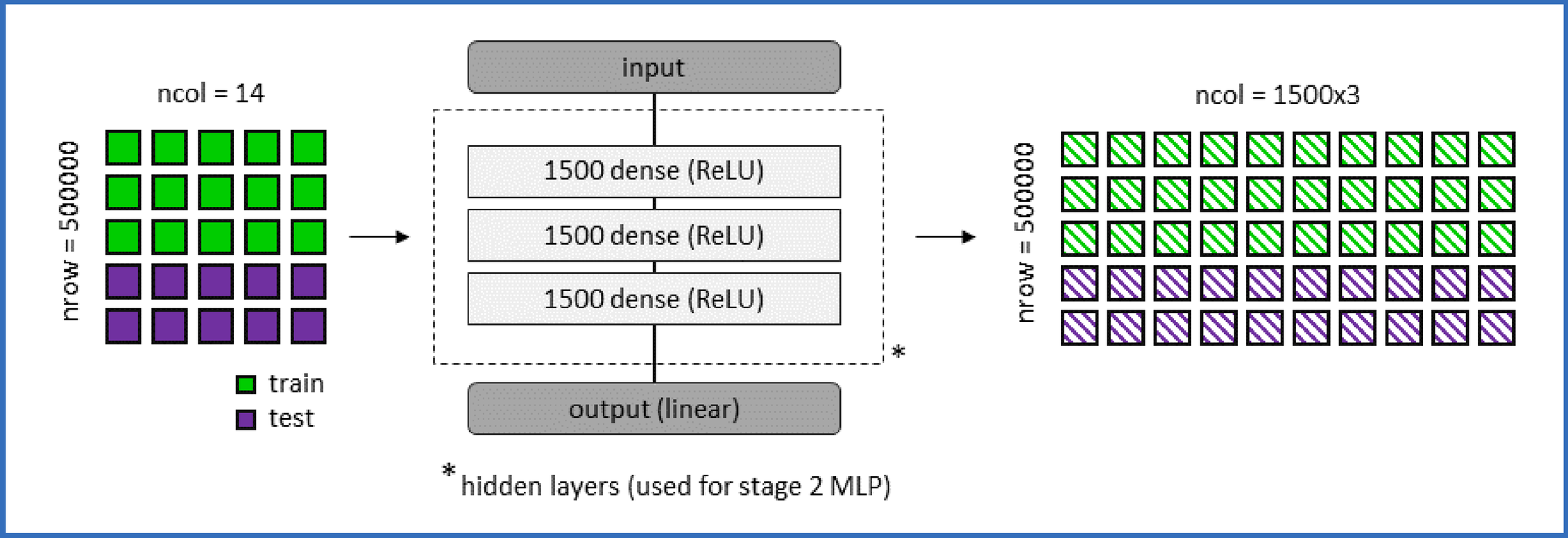


- We also tried refactoring the problem to image classification
- Each row of the table is reshaped into a kernel and the kernel is convoluted on a standard image.
- ResNet-34 is fine tuned to classify the convoluted image.

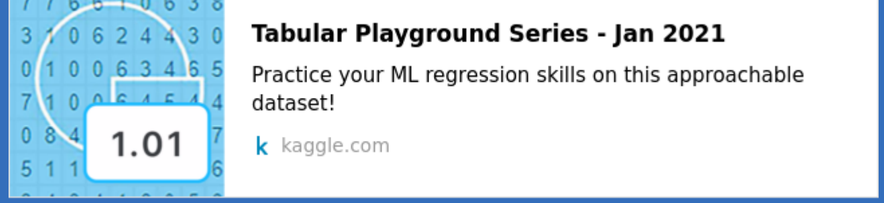
FEATURE EXTRACTION USING SELF-SUPERVISED DENOISING AUTO-ENCODER

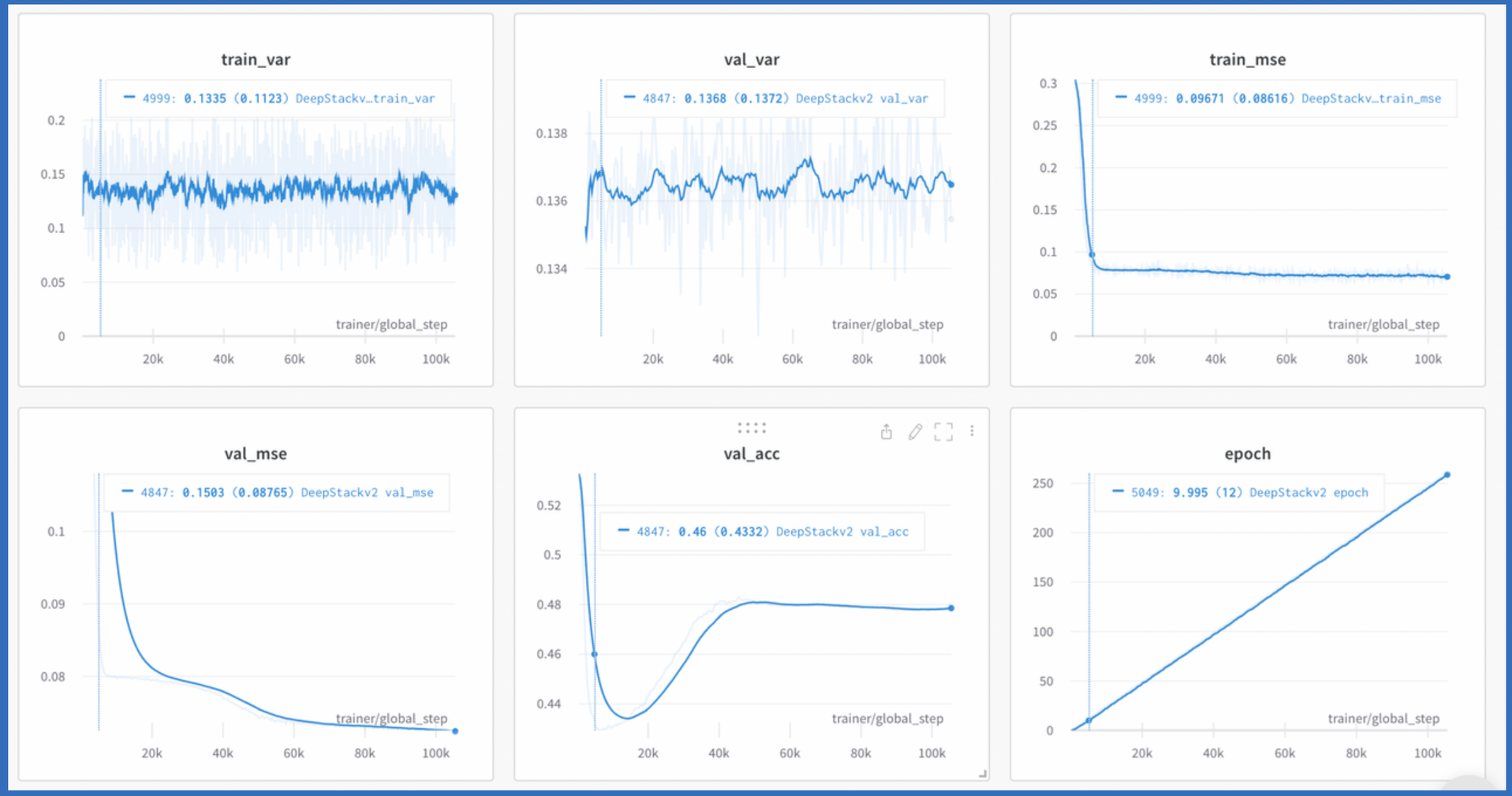


- The core idea behind this auto-encoder is to add noise to tabular data and build a model to predict the original data.
- The layer wise outputs of the model are concatenated to form new feature set for a second model.

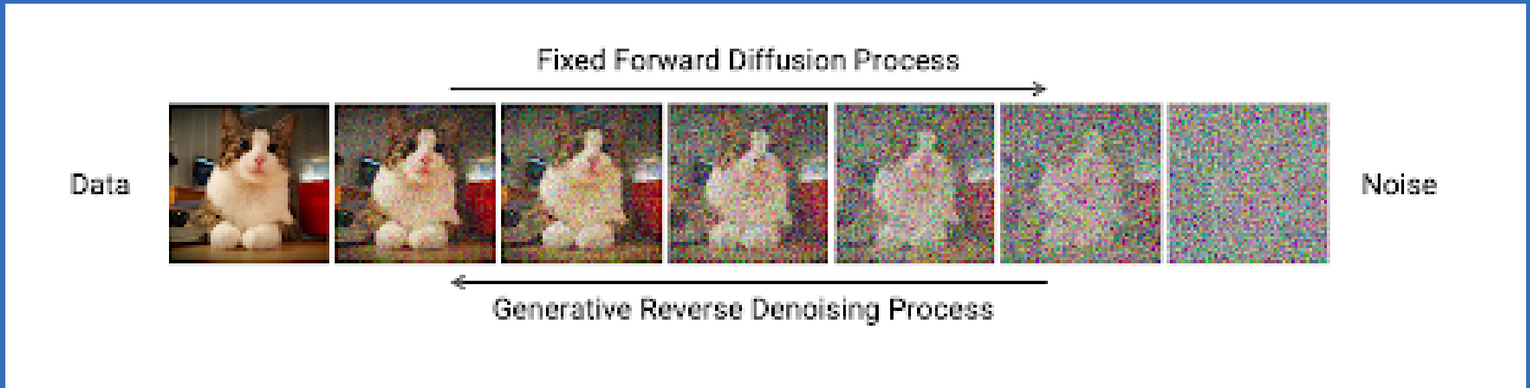


FEATURE EXTRACTION USING SELF-SUPERVISED DENOISING AUTO-ENCODER



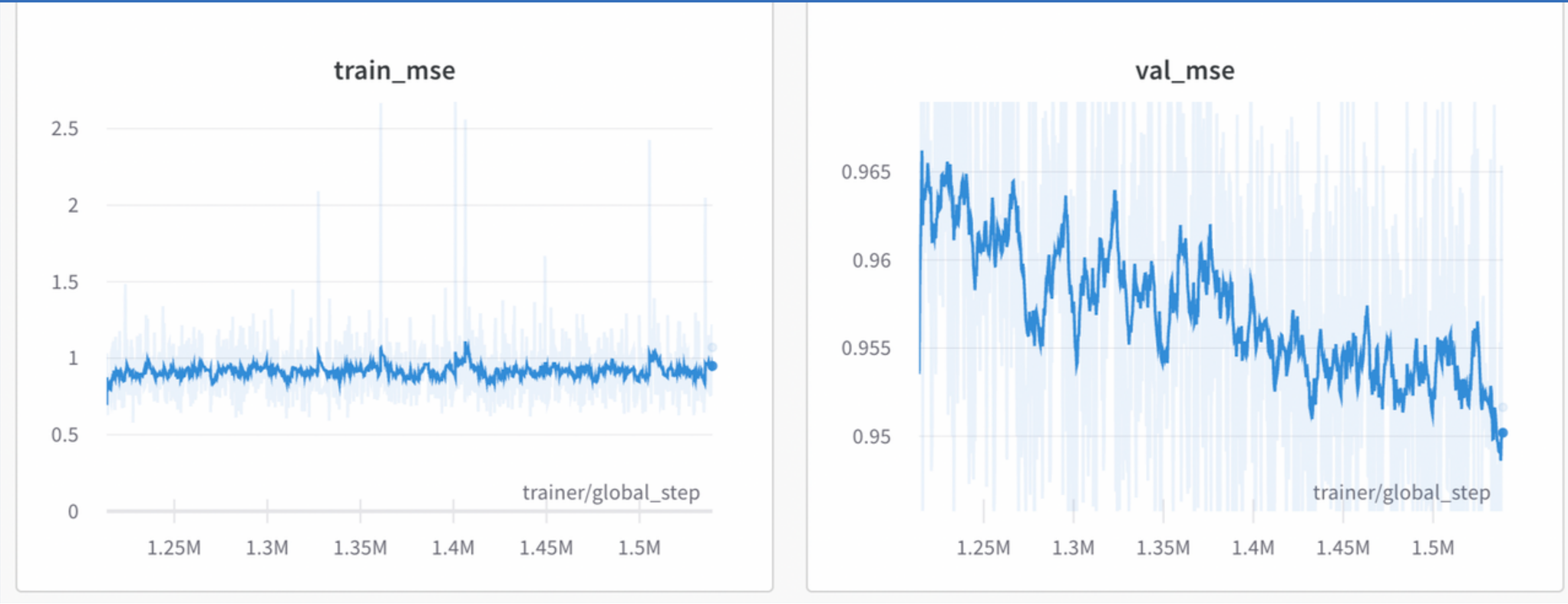


DENOISING DIFFUSION MODEL



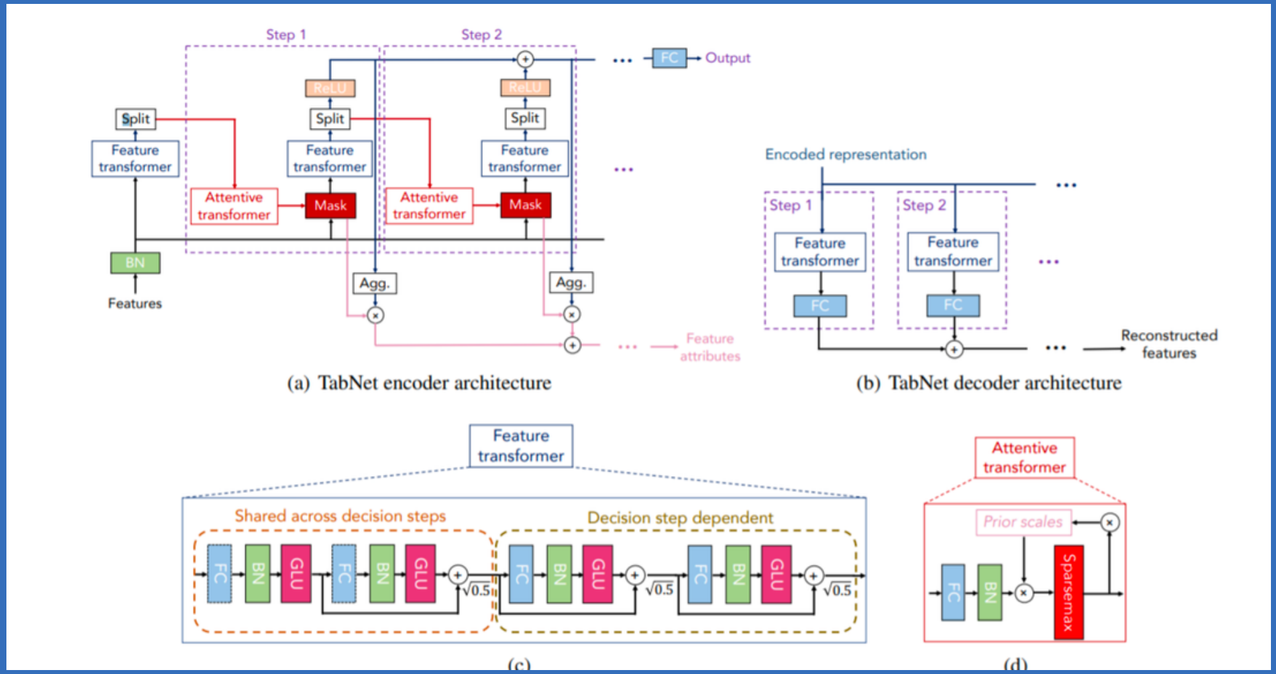
- In recent times, Denoising Diffusion Probabilistic Models have shown some stunning performance in image generation.
- Modern image generation models like DALL-E, Stable Diffusion uses DDPMs for state of the art image generation.
- We implemented 1D version of DDPMs, DDPMs can be used to remove noise in the data as well as generate literally infinite number of data.

DENOISING DIFFUSION MODEL



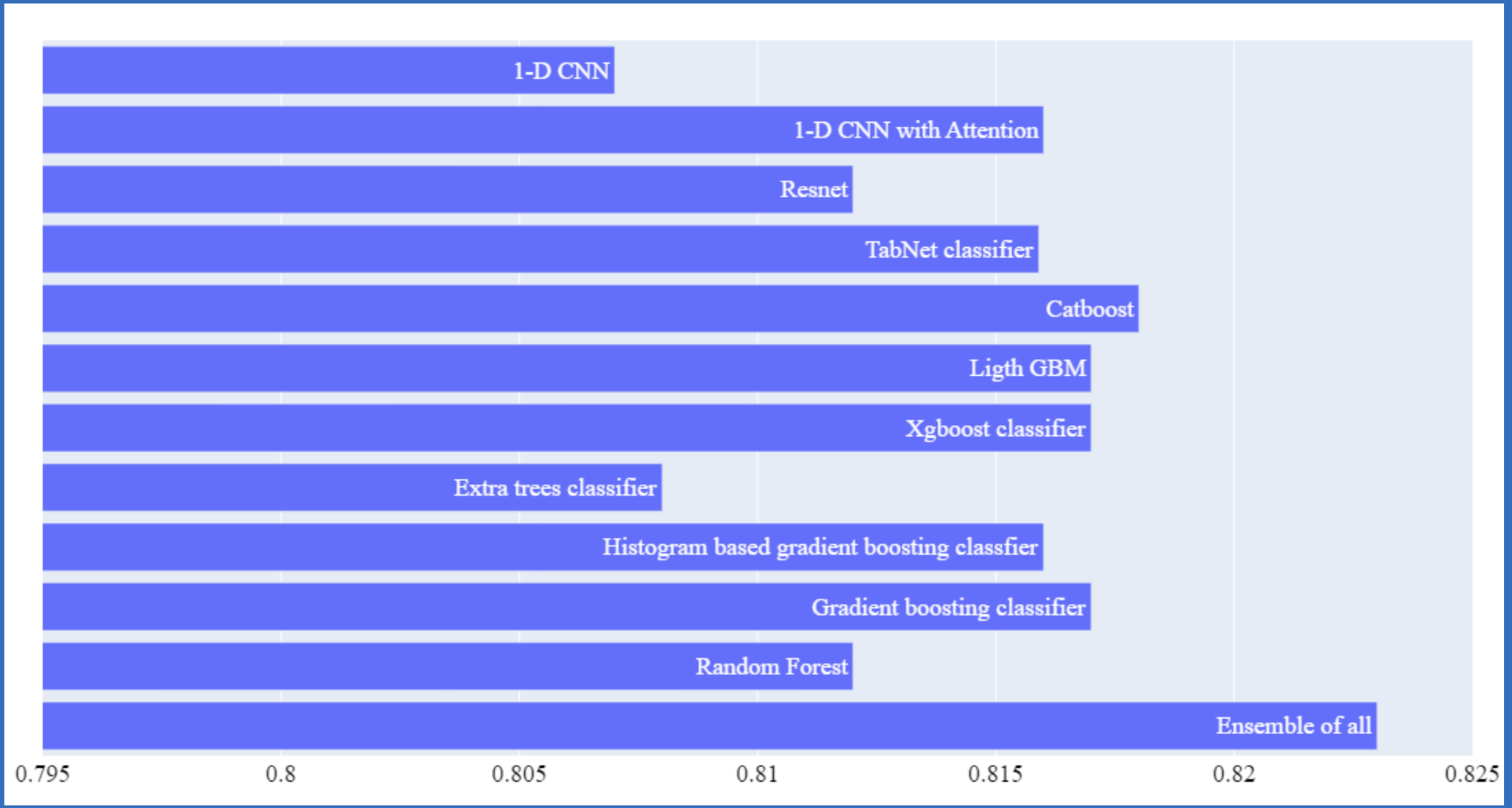
- Very slow training!

TABNET



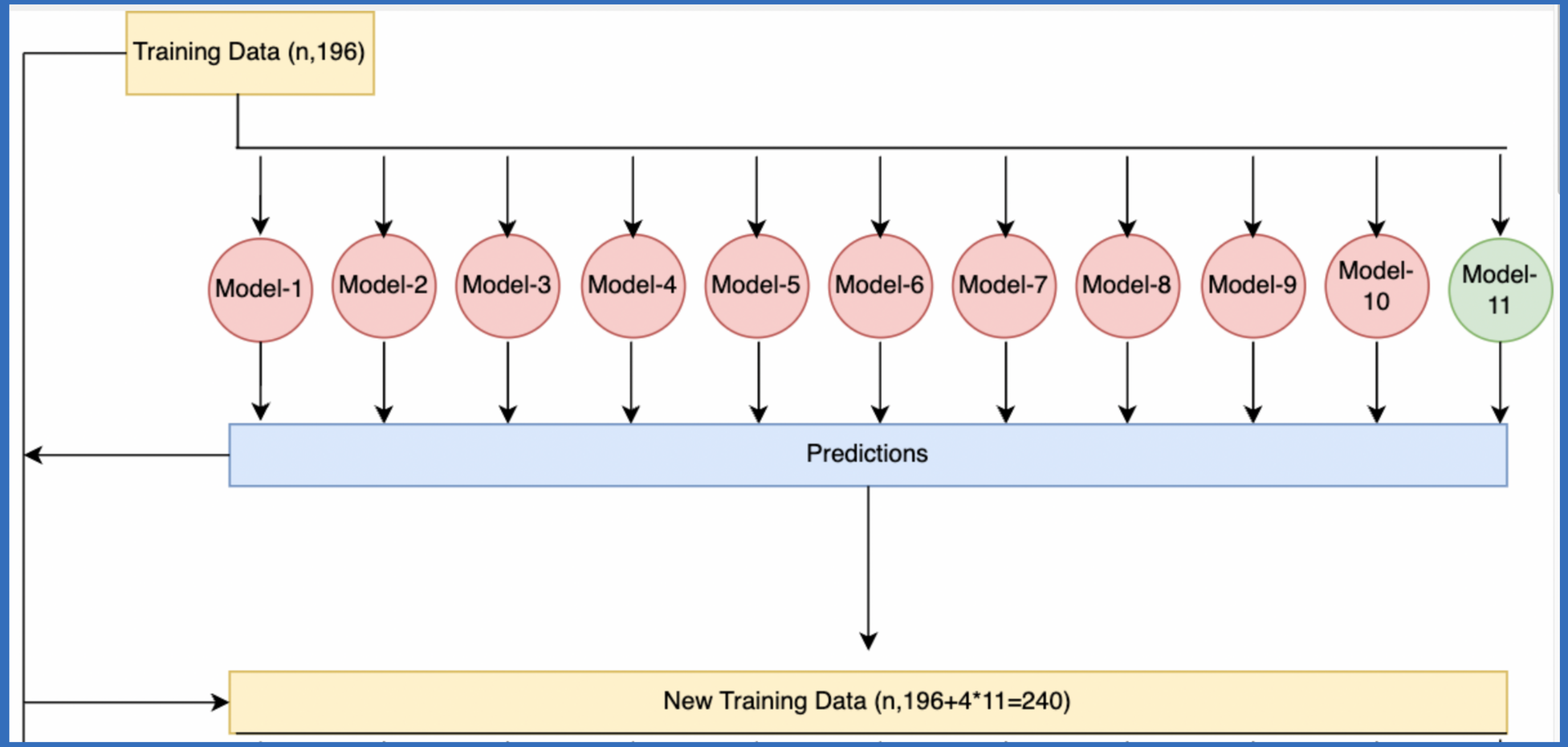
- The core idea behind TabNet is to apply deep neural networks to tabular data.
- TabNet uses a technique known as the sequential attention mechanism

MODELLING


1-D CNN
1-D CNN with Attention
Resnet
TabNet classifier
Catboost
Ligth GBM
Xgboost classifier
Extra trees classifier
Histogram based gradient boosting classfier
Gradient boosting classifier
Random Forest
Ensemble of all
0.795
0.8
0.805
0.81
0.815
0.82
0.825

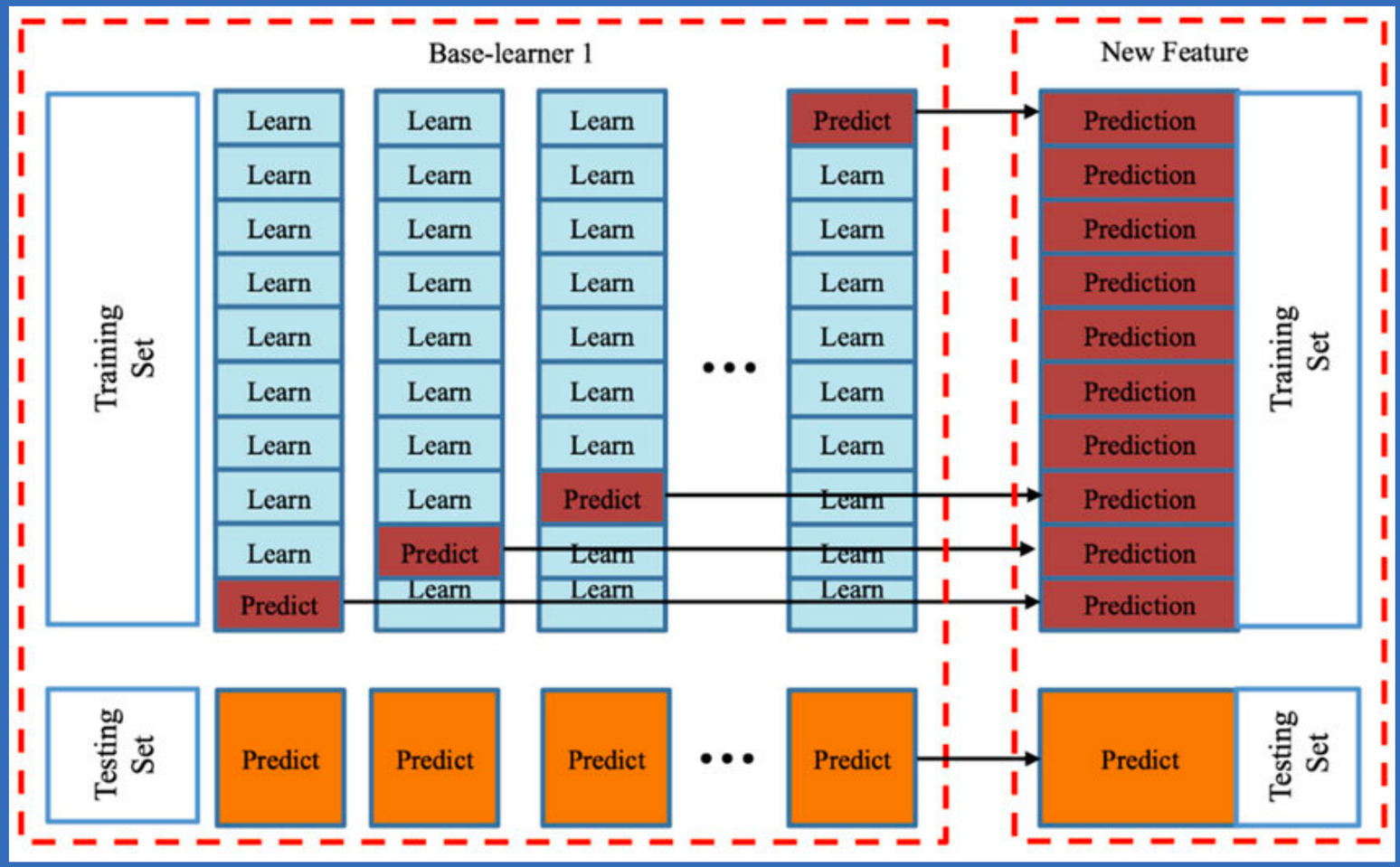
FINAL MODEL

4 LEVEL STACKING WITH 10 FOLD BAGGING



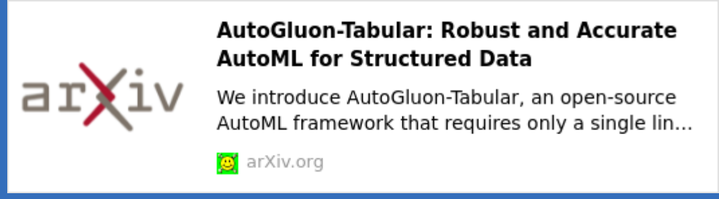
FINAL MODEL

4 LEVEL STACKING WITH 10 FOLD BAGGING



FINAL MODEL

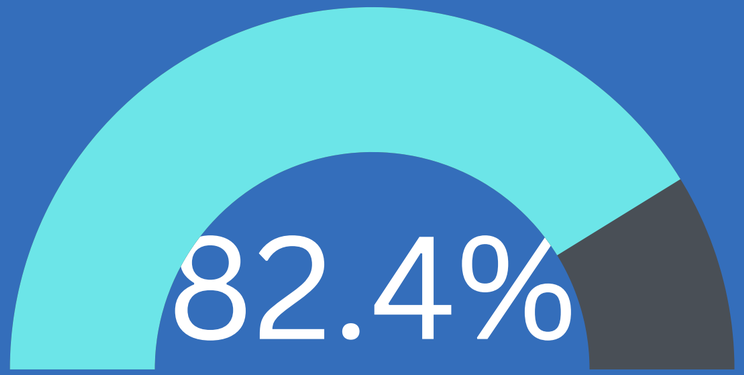
4 LEVEL STACKING WITH 10 FOLD BAGGING

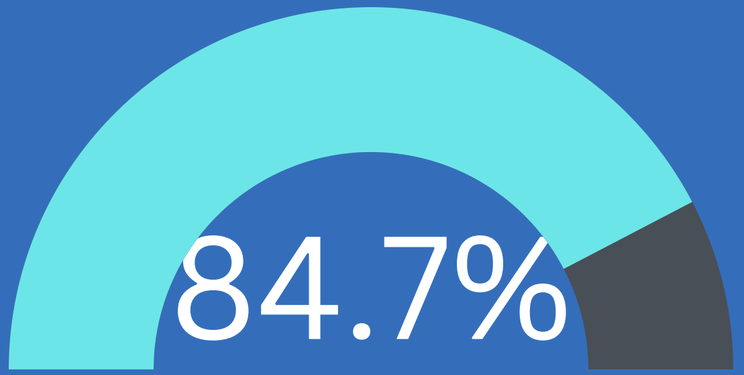


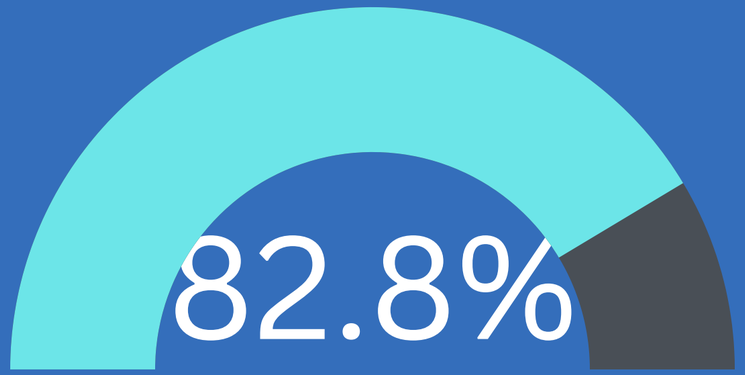
These are the models we used for the final ensemble model

- Random Forest with Entropy as criterion
- Random Forest with Gini impunity as criterion
- Extra Trees with Entropy as criterion
- Extra Trees with Gini impunity as criterion
- LightGBM
- LightGBM with extra trees
- CatBoost
- XGBoost
- Neural-Net-Torch
- Neural-NetFastAI

RESULTS







Accuracy: 0.824

Recall: 0.847

F1 Score: 0.828

THANK YOU