Group1_Phase3_HCDR.

Anitha:

Hello Everyone, Welcome for our presentation.
For Final Phase 3 for HCDR we were asked to build on our work from Phase 2 and build a neural network using PyTorch.

Archana:

EDA Summary
In the Final Phase we went over the EDA to validate the features from our feature importance and found feature related to occupation time, installment annuity and education type categories stood out.

We revisited our feature engineering process to handle multi-collinearity , zero variance and it resulted in a whopping 495 features. We also tried to ensure that our train and test data was separated.

Nisha: End to End ML Pipelines.
End to End ML pipeline & Workflow
We used the end to end pipeline to create HCDR predicting model. In Phase-1, we implemented Logistic regression model as the baseline model with sampled and imbalance data . In phase 2 we explored various classification models and Our primary focus was on boosting algorithms with RFE feature selection algorithm. In our final phase, we experimented the classifiers with various feature selection algorithm like PCA, SelectKBest and Variance Threshold, and SMOTE,Early stopping to avoid data leakage and over fitting. We expanded our project by creating single and a multi-layer deep learning models, including linear, sigmoid, ReLu, and hidden layers. We used binary CXE, custom hinge loss with adam & sgd optimizer.

Rajesh:

Experiment Results + Kaggle submission
We totally performed 18 experiments with 132 features. Our best model turned out to be Logistic Regression with SelectKBest feature selection with 74.86% ROC score. Our hopes were higher on XGBoost classifier with early stopping technique but it stood out to be second best in our models. Our Deep Learning of simple network preformed model better than the multilayer network

with the ROC Score as 74.6% for the simple network. For multilayer network our score came as 59.38%. Compared to traditional machine learning model, training a deep learning model is computationally efficient. The deep learning Kaggle score fell short of the ensemble model.

Anitha:

Problems faced
The problem encountered apart from the accuracy of the model include:
1. Handling large data with too many features
2. Implementation of the neural network.

Additional experimentation will result in a better performing deep learning models. By combining and continuing to refine our extended loss function, we can further demonstrate our effectiveness.
