HCDR TRANSCRIPT PHASE – 2

Chandra Sagar: **Introduction**

1) Hello this is group 9 with the HCDR project, with myself, Hanish, Nidhi and Srinivas
2) The contents are as follows.
3) The four p's in this project are as follows:
a) For our completed phase zero this was all about setting the scope of the project, understanding the project requirements and deadlines. the biggest items were identifying the machine learning algorithms and identifying the loss functions and metrics that we wanted to apply for the project.
b) In the present we did EDA and built a baseline logistic regression pipeline and rebalanced the data and we calculated accuracy and AUC.
c) In the planned phase 2 we plan to use decision making trees, random forests and SVM.
d) The problem we are facing is we need more prior knowledge about the data which will help us in feature engineering and hyperparameter tuning.

Hanish Chidipothu: **Feature Selection**

Feature selection is the process of reducing the number of input variables when developing a predictive model.

- We dropped the irrelevant features and have discarded features with the missing values above 30%
- We filled the missing values like the
  - ❖ CNT Social circle with 0
  - ❖ CNT Fam members with median.
- Related features were added based on priority like the salary to credit.

Nidhi Vraj: **Hyperparameter Tuning**

- After Feature Engineering, we tuned our model to find the optimal parameters. We used Decision Making Tree in addition to the baseline and selected Logistic Regression.
- The hyperparameter tuning for grid search was conducted for decision Making Tree, Lasso Regression, Ridge Regression, XG Boost and Logistic Regression.
- For the Decision Tree model, we used different maximum depths and the number of sample splits.

- For Lasso and Ridge Regression, we used different alpha parameters and controlled the weighting of the penalty to the loss function.
- We used different maximum depths and trees for XG Boost.
- For Logistic Regression, we used different C parameters and controlled the penalty strength.

Srinivas Vaddi: **Conclusion**

As we can see from the results the feature engineering did really help in improving the model's accuracy.

The baseline linear regression's accuracy is a fair bit different from the previous time. Crossfold train accuracy we got 91.8% and more in line with 91.9%. The area under the curve is at 50% now. So is the case with baseline with 79 I/P and Grid search on 79 I/P.

AUC, which is an indicator of performance and model's comprehensiveness, is improved with a decision tree model which has a slightly higher test MSE in contrast to the later ones like L1 and L2 regularizations (Lasso, Ridge Regressions respectively) and XG Boost.

To summarize, several models were estimated, Feature engineering has been done with the idea to improve the models accuracy.

Feature engineering, data imputing and hyper parameter tuning are further steps. We will implement a deep learning model and we will build additional models in pytorch and submit our results in kaggle.