Hello everyone, this is group 11, and we are going to present phase 3 of our hdcr project. This is how our project phase 3 is going to look like in outline.

The goal of our project is to create a model that can forecast if a potential borrower of a new mortgage will be a "default risk" or not.

In phase 1 and 2 we completed data preparation, performed EDA and built baseline models.

Last time only some polynomial feature engineering. But now along with that we took two more tables bureau.csv and previous\_ application.csv. We performed aggregate feature engineering and merged those based on customer id that are preset in previous application and application training

Coming back to the hyper parameter tuning we sub sampled the entire dataset into 2 ways: into 5000 rows and 10000 rows using stratified strategy. We used the following classifier such as gradient boosting, random forest, XGBoost, logistic regression and AdaBoost. We also used ensemble and non ensemble classifier. On evaluating all the models on train and test AUC scores we came to know that Adaboost was the best performing model. We also performed grid search and random search. In grid search results were slightly better than random search. The mode was slightly over fitting for random forest and decision trees.

After submitting to Kaggle, we received an AUC private score of 70.4% and a public score of 71.6%.

Based on the results that we just presented we have concluded that our best performing model is logistic reg and ADAboost classifier.

Talking about the four P’s of this phase. These are the past present and future work done or to be done by the team and the problems we faced during this phase.